计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(1):29−36 [doi: 10.15888/j.cnki.csa.008282]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于贝叶斯网络的食品安全舆情监控探针研究①

王　旎[1,2], 孙晓红[1], 吴　锴[3], 谢　锋[2,3], 陶光灿[1,2,3]

[1](贵州医科大学 公共卫生学院, 贵阳 550025)
[2](贵州省分析测试研究院, 贵阳 550014)
[3](食品安全与营养 (贵州) 信息科技有限公司, 贵阳 550014)
通信作者: 陶光灿, E-mail: tgcan@gzata.cn

摘　要: 针对大数据时代食品安全舆情数据采集不够快捷与准确的问题, 提出一种基于贝叶斯网络的食品安全舆情监控探针的研究方法. 首先, 通过 MySQL 数据库建立食品安全关键词库; 其次, 运用贝叶斯网络模型将关键词库构建形成监控探针, 并选定人民众云舆情监测系统进行数据采集; 第三, 将监控探针与传统舆情数据采集、网络爬虫技术做 3 组对比实验 (奶类、酒类、茶类), 验证其有效性. 结果显示 3 组实验的数据挖掘时间 (乳制品类 3 s; 酒类 2.5 s; 茶类 2.4 s) 明显降低, 数据有效率 (乳制品类 83.6%; 酒类 77%; 茶类 77.9%) 明显升高. 可见关键词库引入贝叶斯网络模型形成监控探针, 可有效提高食品安全舆情数据采集的及时性与精准度.

关键词: 食品安全; 舆情监测; 数据采集; 贝叶斯网络; 监控探针

引用格式: 王旎,孙晓红,吴锴,谢锋,陶光灿.基于贝叶斯网络的食品安全舆情监控探针研究.计算机系统应用,2022,31(1):29–36. http://www.c-s-a.org.cn/1003-3254/8282.html

Research on Public Opinion Monitoring Probe on Food Safety Based on Bayesian Network

WANG Ni[1,2], SUN Xiao-Hong[1], WU Kai[3], XIE Feng[2,3], TAO Guang-Can[1,2,3]

[1](School of Public Health, Guizhou Medical University, Guiyang 550025, China)
[2](Guizhou Academy of Testing and Analysis, Guiyang 550014, China)
[3](Food Safety and Nutrition (Guizhou) Information Technology Co. Ltd., Guiyang 550014, China)

Abstract: To address the problem that the public opinion data collection on food safety is not fast and accurate enough in the era of big data, this study proposes a public opinion monitoring probe on food safety based on the Bayesian network. Firstly, the MySQL database is used to establish a food safety keyword database. Secondly, the Bayesian network model is adopted to build a monitoring probe with the keyword database, and the public opinion monitoring system of the "Zhongyun Big Data" of PeopleYun is chosen for data collection. Thirdly, the monitoring probe is compared with traditional data collection technologies on public opinions and Web crawler technologies in three groups of comparative experiments (milk, wine, and tea) to verify its effectiveness. The results show that the data mining time of the three groups of experiments (milk: 3 s; alcohol: 2.5 s; tea: 2.4 s) is significantly reduced, and the data efficiency (milk: 83.6%, alcohol: 77%, tea: 77.9%) is considerably enhanced. Therefore, introducing a keyword database into the bayesian network model to form a monitoring probe can effectively improve the timeliness and accuracy of public opinion data collection on food safety.

Key words: food safety; public opinion monitoring; data acquisition; Bayesian network; monitoring probe

① 基金项目: 国家重点研发计划 (2017YFC1601800); 贵州省科技计划 (黔科合平台人才 [2018]5404)
收稿时间: 2021-03-22; 修改时间: 2021-04-19, 2021-05-07; 采用时间: 2021-05-11; csa 在线出版时间: 2021-12-17

在网络新媒体时代, 为促进食品行业健康发展, 食品安全网络舆情监测体系应运而生, 开发决策参考、监督抽检、专项整治、协查处置等舆情信息应用场景化服务, 针对当下热门的食品安全舆情事件自动展开跟踪与分析[1]. 而数据采集作为舆情大数据资源池建设的第一步准备工作, 将孤立分布在数据报刊、网络媒体、微博、微信中的各个数据源采集并存储, 为下一步舆情分析打下基础, 帮助政府、企业和舆情相关者采取措施以预警或控制食品安全舆情的发展态势[2]. 但是, 食品安全舆情数据采集在及时性和精准性等方面仍存在着许多痛点和难点问题[3], 一方面, 运用传统语义识别的方法采集数据所需的费用偏高且准确率较低, 采集内容要素广泛且难以统一, 包括食品类别、风险类型、健康危害等多种关键词, 数据报刊、网络媒体等多个舆情渠道, 食品企业、政府和消费者等多方面用户对象, 以及包括监督抽检、检测机构、急救中心、公安部门的其他关联因素. 在采集过程中, 通过传统人工采集数据的方式难以穷尽, 新发生的食品安全事件都有新的关键词; 另一方面, 由于数据采集不够精准, 无法有效减少垃圾数据的产生, 必然会影响数据采集的效率.

贝叶斯定理是数据挖掘领域一种用来描述概率关系的算法[4], 提出了将知识图解可视化的推理和模型[5], 其方法简单、分类准确率高、速度快, 模型参数估计不需要任何复杂的迭代求解公式, 只需统计训练集中的先验概率和条件概率[6]. 目前已广泛应用于医疗诊断、人工智能、生物信息学、金融分析与预测等多个领域[7]. 因此, 本文运用贝叶斯网络模型优化食品安全关键词库的风险概率, 将高风险性的食品类别、风险类型和健康危害等输出为关键词组, 做一个自动关联风险的数据模型, 并形成监控探针, 结合舆情监测系统, 实现采集食品安全舆情关键词的合理配置, 以提高采集效率和准确率. 其中, 监控探针[8] 是一个不流行的学术术语, 常用于描述语言及其编译器的设计, 对其功能阐述为嵌入在目标系统代码中, 在系统运行时获取目标监控点的相关运行状态.

基于以上研究现况, 提出科学假设: 构建关键词库形成基于贝叶斯网络的监控探针, 可提高食品安全舆情数据采集的及时性与精准度. 对比实验: 运用传统人为设计关键词、网络爬虫和监控探针的 3 种方法采集同一食品安全事件的舆情数据, 从而对监控探针设计的快捷性和准确性进行验证.

1 现有工作

目前, 网络舆情数据采集的框架主要由 6 部分组成: 网站页面、链接抽取、链接过滤、内容抽取、网络爬虫技术[9] 和数据[10]. 其中, 新时代背景下又增加了爬行策略设计、网页更新策略、网页去重和计算机转换软件等新兴互联网采集技术, 针对结构化、半结构化和非结构化的网络数据进行汇总和收集[11]. 在算法上, 主要采用分布式、并行式的计算模型, 以提高数据采集的速度[12]. 在信息存储技术上, 主要用 Oracle、MySQL 数据库和 HBase、MongoDB 数据库来实现[13]. 从应用范围上, 网络爬虫技术[14] 和信息抽取技术[15] 是目前主流的舆情数据采集和分析挖掘方法. 网络爬虫最早应用于搜索引擎中, 用来收集媒体网页中的数据, 抓取有效舆情信息并加以存储[16]. 信息抽取技术 (information extraction, IE) 是从非结构化的自然语言文本中提取目标信息, 然后进一步转换成结构化数据形式的采集方法[17]. 伴随科技的发展, 全文信息的搜索引擎逐渐不能很好地满足用户要求, 1957 年, Luhn[18] 提出一种基于词频统计的关键词抽取方法, 衍生出一系列关键词抽取技术[14]. 关键词抽取分为: (1) 基于统计的方法, 该方法的主要思想是通过指定特征来对词语的权重进行计算, 并根据词语的权重大小来抽取关键词[19]. 例如频率统计 (TF-IDF)[20] 及其改进方法, 简单易行, 具有较强的适用性, 但由于 TF-IDF 只提取频率较高或位置较特殊的关键词, 不能完整概括全文主体信息, 导致数据采集的准确率降低[21]. (2) 基于语言规则的方法, 通过从文章、句子以及词语等层次进行语法分析, 来提高关键词抽取系统的性能. (3) 基于人工智能的方法, 让计算机能够自动学习关键词抽取的过程, 通过对模型进行训练以实现人工智能自动抽取关键词.

但是, 互联网数据具有海量、异构数据源、缺乏语义信息和动态可变性等特点, 尤其是网络社交媒体和新闻数据, 其更新频率高, 随时随地都会生产出大量信息, 这部分内容数据量大, 交互性强, 使得抽取技术变得更加复杂化, 给舆情信息的抽取带来了诸多困难. 并且, 目前在网络上针对食品安全的舆情数据并没有系统地进行过汇总, 采集者缺乏食品安全专业领域的知识, 使得无论是应用网络爬虫技术还是信息 (关键

词) 抽取技术都没有高效的关键词以供参考, 不仅无法精准定位采集对象, 还浪费了舆情数据的挖掘时间, 造成了大量垃圾数据的产生, 增加了舆情数据采集的工作量和计算成本, 影响了舆情监测系统的运营进度, 阻碍了食品安全舆情监测系统的拓展应用.

因此, 针对大数据时代食品安全舆情数据采集不够快捷与准确的问题, 从采集关键词的研究对象出发, 构建有关食品安全的关键词库, 并引进贝叶斯网络模型的分析方法, 将高风险的食品安全关键词设计成监控探针, 向舆情监测者推荐采集较高的关键词组, 提高食品安全舆情数据采集的速度与效率, 对于改善食品安全舆情监测系统的数据采集环节, 为后续的数据处理与数据应用打下良好的基础建设, 提高国家在食品安全舆情方面的管理能力具有重要作用.

贝叶斯网络 (Bayesian network, BN)[22] 将贝叶斯理论、图论、人工智能和决策分析相结合, 是一种基于概率推理的图形化网络属性数学模型[23]. 其在态势评估、医疗保健、工业风险预测等领域都已有广泛的应用, 比如, 根据环境、人员等因素引入贝叶斯网络对采矿现场、建筑施工等高危作业是否发生事故进行风险预估, 以减少风险事件的发生[24]; 为提高疾病诊断效率, 提出了基于余弦相似度加权改进的贝叶斯分类算法[25], 开发了大量的人工智能方法辅助检测疾病[26]; 在科技发展与人类智能上, 提出了基于贝叶斯理论的人脸识别算法[27]; 基于贝叶斯算法的垃圾邮件识别与过滤系统[28]; 以及基于贝叶斯网络的民航机票预测系统等[29].

2 材料与方法

2.1 数据来源

北京人民在线网络科技有限公司的人民众云平台[30].

烟台富美特信息科技股份有限公司食品伙伴网的食品标准 (国家标准)[31].

食品安全与营养 (贵州) 信息科技有限公司食品安全云平台的技术社区 (国家标准)[32].

2.2 食品安全舆情监控探针总体框架

首先, 将食品类别、风险类型、健康危害等食品安全舆情关键词, 利用 MySQL 数据库构建形成统一完善的食品安全关键词库; 然后, 运用贝叶斯网络算法对关键词库建立数学模型, 得出高风险性的食品安全舆情关键词组, 并自动关联风险形成 监控探针; 最后, 结合北京人民在线网络科技有限公司开发的食品安全舆情监测系统, 向用户提供优先采集的关键词选项, 形成一个包含监控系统、关键词库、数学模型的监控探针, 从而达到提升食品安全舆情数据采集及时性与精准度的效果, 如图 1 所示.

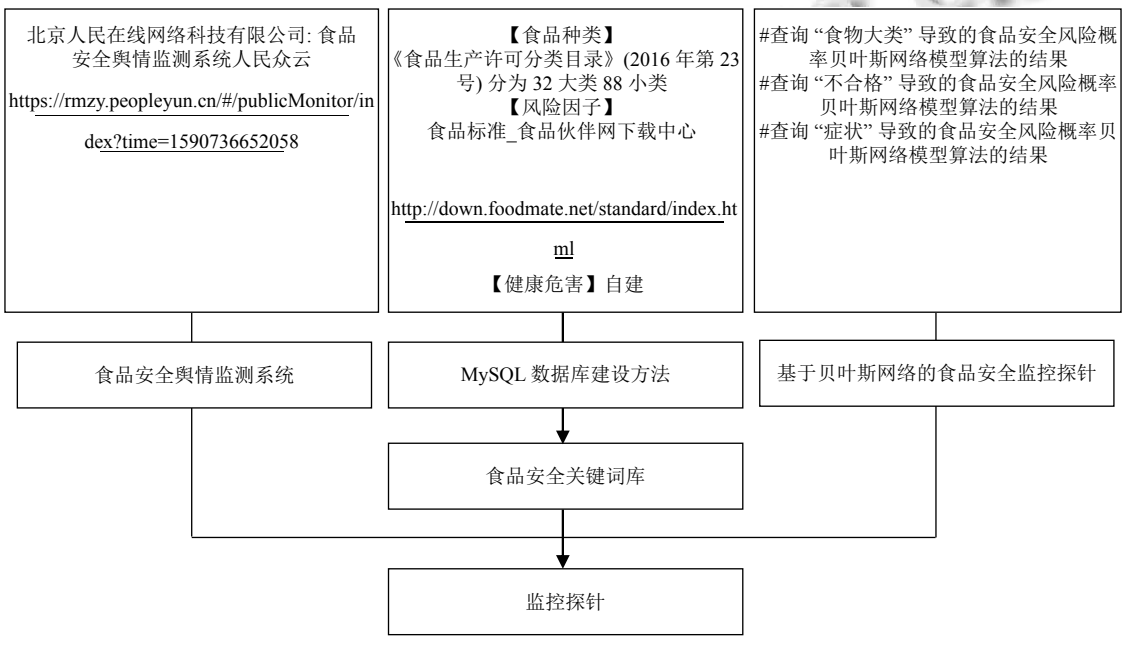

图 1 食品安全舆情监控探针逻辑图

2.3 构建关键词词库

为了更准确描述一个食品安全事件的语义模板, 包含发生地域、食品类别、风险因子以及造成的健康危害等关键词, 构建食品安全舆情事件信息关键词分类表 (见表 1), 并做出以下定义: 定义 1. 设 a 为食品安全舆情事件发生地域关键词, 地域分布以省、直辖市、自治区为父类, 下辖地级市为子类, 共计 34 个省级行政区; 定义 2. 设 b 为食品安全舆情事件谓语表达关键词; 定义 3. 设 c 为食品安全舆情事件食品类别关键词, 食品分类方法以国家市场监督管理总局颁布的《食品生产许可分类目录》[33] 为依据, 共计 32 类; 定义 4. 设 d 为食品安全舆情事件风险因子关键词, 风险因子指能够促使或引发食品风险事件的危害要素, 分为生物性因素、化学性因素、物理性因素和人为因素等[34]; 定义 5. 设 e 为食品安全舆情事件健康危害关键词, 即风险因子可能导致的人体健康损害.

表 1 食品安全舆情事件信息关键词分类

空间分布a	谓语表达b	食品类别c	风险因子d	健康危害e
黑龙江、北京、上海、湖北、天津、浙江、重庆、河北、江苏、山西、陕西、山东、河南、辽宁、吉林、安徽、江西、福建、湖南、贵州、四川、云南、海南、广东、甘肃、青海、内蒙古、新疆、广西、宁夏、台湾、香港、澳门、西藏	报道、曝光、举报、发现、检验通报、检出、存在、公布造假、假冒、掺水等	食用油、油脂及其制品, 粮食加工品, 食品添加剂, 肉制品, 调味品, 乳制品, 方便食品, 饮料, 饼干, 冷冻食品, 罐头, 速冻食品, 糖果制品, 薯类和膨化食品, 茶叶及相关制品, 蔬菜制品, 酒类, 水果制品, 蛋制品, 炒货食品及坚果制品, 可可及焙炒咖啡产品, 淀粉及淀粉制品, 食糖, 水产制品, 糕点食品, 蜂产品, 豆制品, 保健食品, 婴幼儿配方食品, 特殊医学用途配方食品, 特殊膳食食品, 其他食品	变质, 菌落超标, 病原微生物污染, 致病微生物污染, 真菌毒素污染, 霉菌超标, 大肠杆菌超标, 寄生虫等生物性, 农药残留, 兽药残留, 食品添加剂, 使用违禁渔药, 肉毒毒素, 黄曲霉毒素B1, 马铃薯毒素, 氰化物, 酸价, 过氧化值, 吊白块, 亚硝酸盐, 重金属等化学性, 包装漏气, 包装密封性欠佳, 超过保质期, 贮存不当, 蟑螂, 苍蝇等物理性, 非法回收, 无健康证, 无证, 无照, 不清洁, 不消毒, 原材料筛选不严格, 更改生产日期等人为性风险因子	神经系统和消化系统的损伤: 肌肉振颤, 心慌, 战栗, 头疼, 恶心, 呕吐, 四肢无力, 休克, 腹痛, 腹泻, 肝肾功能、泌尿系统损伤: 膀胱、肾结石, 金属中毒, 食物中毒, 肥胖, 龋齿, 骨质疏松, 消化不良, 致癌, 致畸

定义 6. 满足食品安全舆情事件条件下, a, c 之间存在谓语 b, 且 c 后为风险因子 d, 造成影响 e, 则称“a, b, c, d, e”5 个词组成一个食品安全舆情事件的标准语义模板. 示例: 2014 年 7 月 20 日东方卫视报道: 上海福寿喜集团存在大量采用变质肉原料的行为, 引发顾客的食物中毒, “上海, 报道, 肉原料, 变质, 食物中毒”对应“a, b, c, d, e”是满足食品安全舆情事件的语义模板 (见表 2). 根据标准语义模板中的语义信息量, 定义了一、二、三、四、五级语义模板. 由此得出, 一件食品安全事件的关键词越齐全, 事件描述越完整, 挖掘到的食品安全信息便越丰富, 对于舆情数据采集工作的意义越大.

表 2 食品安全舆情事件多级语义模板

级别	一级语义模板	二级语义模板	三级语义模板	四级语义模板	五级语义模板
描述信息量	完整描述食品安全事件	较完整描述食品安全事件	少量描述食品安全事件	仅描述两类食品安全事件关键词	仅描述一类食品安全事件关键词
标准语义模板	“a, b, c, d, e”	“a, b, c, d”	“a, b, c”	“a, b”	“a”或“b”
食品安全事件示例	上海市报道福寿喜集团使用变质肉原料致顾客食物中毒	天津高女士举报某菜市猪肉存在变质问题	2006年北京爆出苏丹红鸭蛋事件	湖北食品添加剂超标	成都七中食堂卫生不合格事件

2.4 使用食品安全舆情监测系统采集数据

首先, 登录食品安全舆情监测系统“人民众云[30]”用户端, 选择“自助监测”栏目, 进行食品安全舆情数据采集的任务设置, 对任务名称和选择分组的基本信息进行填写; 然后, 进入“关键词选择”功能, 填写“主关键词”“辅关键词一”“辅关键词二”“辅关键词三”, 并设置数据采集范围: “数据报刊、政府机构、网络媒体、网络视频、微博、微信、资讯、论坛等”; 最后, 提交操作采集舆情数据.

2.5 运用 MySQL 数据库建立食品安全关键词库

MySQL 数据库是一种高速度、高性能、多线程、开放源代码的关系型数据库管理系统[35], 是互联网行业存储和操作数据最常用的数据库[36]. 根据贝叶斯网络节点进行设计数据存储, 每条数据包括食物大

类 (*F* 节点)、风险因子 (*R* 节点)、危害症状 (*S* 节点)3 个关键词, 分为 32 个食品类别[28], 其中食品添加剂 18 条、粮食加工品的关键词条 42 条、食用油 30 条、调味品 52 条、乳制品 36 条、饮料 60 条、方便食品 40 条、肉制品 77 条、饼干 44 条、冷冻饮品 22 条、罐头 33 条、速冻食品 12 条、糖果制品 7 条、薯类和膨化食品 54 条、茶叶及其制品 4 条、蔬菜制品 25 条、酒类 30 条、水果制品 25 条、糕点食品 36 条、蛋制品 18 条、可可及焙烤咖啡产品 49 条、炒货食品及坚果制品 84 条、水产制品 60 条、淀粉及淀粉制品 30 条、蜂制品 56 条、豆制品 12 条、保健食品 40 条、食糖 5 条、特殊医学用途配方食品 1 条、其他食品 1 条, 共 1 039 条数据.

以“粮食加工品”为例, 在 MySQL 食品安全关键词库中查询“粮食加工品”, 输入查询编程:

```
SELECT a.f0, a.f1, b.f2, c.f3 from testlv1 a
LEFT JOIN testlv2 b on a.f1=b.f1
LEFT JOIN testlv3 c on a.f1=c.f1
WHERE a.f1='粮食加工品'
ORDER BY f0, f2, f3 ASC
```

可得出包括“粮食加工品”食物大类、风险因子、可能症状 3 种关键词的 42 条数据. 每一条数据都具有唯一性, 为食品安全舆情数据采集提供专业性较高的关键词, 提高数据采集的精准度, 减少采集时垃圾数据的产生.

2.6 构建基于贝叶斯网络模型的食品安全舆情监控探针

设置“食品安全风险因子、食品类别、食品检测不合格、食品危害症状”的 4 个变量为贝叶斯网络模型的节点, 确定节点之后, 采用因果推理形式的方法, 确定各节点之间的关系, 由原因推知结果, 以求得食品安全变量导致的风险事件发生的概率, 从而建立有向无环图, 如图 2 所示. 其中, *R* 节点为风险因子 (risk), *F* 节点为食物大类 (food), *S* 节点为症状 (symptoms), *O* 节点为检测不合格 (out of specification, OOS).

根据概率乘法公式有 $P(X)=P(X_i/X_1, X_2, \cdots, X_{i-1})$ 用 P_{ai} 表示变量 X_i 的父节点集, 则 $P(X)=P(X_i/P_{ai})$, 因此为了确定贝叶斯网络结构, 需要: ① 将变量 $X_1, X_2, \cdots, X_i$ 按某种次序排序; ② 确定满足 $P(X)=P(X_i/P_{ai})$ 的父节点集合 P_{ai} (i=1, 2, $\cdots$, n); ③ 指定局部概率分布 $P(X_i/P_{ai})$. 从图 2 可以清楚地看到影响食品安全舆情数据采集的风险节点及其相互的节点关系. 在因果推理中, 当食品安全风险等级为 R=1 时, 概率关系组合如下:

$$\begin{cases} P(R=1/F=1,S=1,O=1) \\ P(R=1/F=2,S=1,O=1) \\ P(R=1/F=1,S=2,O=1) \\ P(R=1/F=1,S=1,O=2) \\ P(R=1/F=2,S=2,O=1) \\ \vdots \\ P(R=1/F=3,S=3,O=3) \end{cases} \tag{1}$$

当食品安全风险等级为 R=2 时, 概率关系组合如下:

$$\begin{cases} P(R=2/F=1,S=1,O=1) \\ P(R=2/F=2,S=1,O=1) \\ P(R=2/F=1,S=2,O=1) \\ P(R=2/F=1,S=1,O=2) \\ \vdots \\ P(R=2/F=3,S=3,O=3) \end{cases} \tag{2}$$

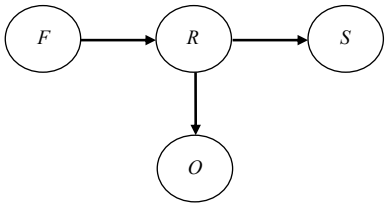

图 2 基于贝叶斯网络的监控探针有向无环图

当食品安全风险等级为 R=3 时, 概率关系组合如下:

$$\begin{cases} P(R=3/F=1,S=1,O=1) \\ P(R=3/F=2,S=1,O=1) \\ P(R=3/F=1,S=2,O=1) \\ P(R=3/F=1,S=1,O=2) \\ \vdots \\ P(R=3/F=3,S=3,O=3) \end{cases} \tag{3}$$

因果推理推出食品安全风险概率:

当 R=3, F=3 时的概率是:

$$P(F=3/R=3)=\frac{P(F=3,R=3)}{P(R=3)} \tag{4}$$

当 R=3, S=3 时的概率是:

$$P(S=3/R=3)=\frac{P(S=3,R=3)}{P(R=3)} \tag{5}$$

当 R=3, O=3 时的概率是:

$$P(O=3/R=3)=\frac{P(O=3,R=3)}{P(R=3)} \tag{6}$$

将贝叶斯网络模型分成食物大类 (*F*)、危害症状 (*S*)、检测不合格 (*O*)3 个互不重叠的部分, 每个部分都可能引发食品安全风险, 且风险发生概率分别为 $P(F)$、$P(S)$、$P(O)$, 引起食品安全风险 J 的可能性就为 $P(J/F)$,

$P(J/S)$, $P(J/O)$. 基于贝叶斯模型的网络风险评估算法:

如果发生食品安全风险, 由“食物大类 F”引起风险的可能性为:

$$P_1 = P(F/J) = \frac{P(F \cap J)}{P(J)} \tag{7}$$

如果发生食品安全风险, 由“症状 S”引起风险的可能性为:

$$P_2 = P(S/J) = \frac{P(S \cap J)}{P(J)} \tag{8}$$

如果发生食品安全风险, 由“检测不合格 O”引起风险的可能性为:

$$P_3 = P(O/J) = \frac{P(O \cap J)}{P(J)} \tag{9}$$

```
#基于贝叶斯网络模型的食品安全风险概率算法
def convert_pgm_to_pgmpy(pgm):
edges=[(edge.node1.name, edge.node2.name) for edge in pgm._edges]
model=BayesianModel(edges)
return model
#定义节点
frisk_model=BayesianModel([('F', 'R'),
                           ('O', 'R'),
                           ('R', 'S')])
cpd_f=TabularCPD(variable='F', variable_card=2,
                 values=[[0.9], [0.1]])
cpd_o=TabularCPD(variable='O', variable_card=2,
                 values=[[0.3], [0.7]])
cpd_r=TabularCPD(variable='R', variable_card=2,
                 values=[[0.03, 0.05, 0.001, 0.02],
                         [0.97, 0.95, 0.999, 0.98]],
                 evidence=['O', 'F'],
                 evidence_card=[2, 2])
cpd_s=TabularCPD(variable='S', variable_card=2,
                 values=[[0.65, 0.3], [0.35, 0.7]],
                   evidence=['R'],
          evidence_card=[2])
#部署模型
frisk_model.add_cpds(cpd_f, cpd_o, cpd_r, cpd_s)
frisk_model.check_model()
infer=VariableElimination(frisk_model)
#查询“食物大类”导致的食品安全风险概率贝叶斯网络模型算法的结果
result=infer.query(['R'], evidence={'F': 1, 'O': 0})
#查询“不合格”导致的食品安全风险概率贝叶斯网络模型算法的结果
result=infer.query(['R'], evidence={'O': 1, 'O': 0})
#查询“症状”导致的食品安全风险概率贝叶斯网络模型算法的结果
result=infer.query(['R'], evidence={'S': 1, 'O': 0})
```

根据可能性大小, 将“风险因子、食物大类、危害症状、检测不合格”4 个方面的关键词设计成监控探针, 按照引起风险的可能性大小, 对高风险性词语实现优先采集, 以提高食品安全舆情数据采集的及时性和精准度.

3 实验结果与分析

运用传统人为设计关键词、网络爬虫和监控探针的 3 种方法采集同一食品安全事件的舆情数据, 针对采集的快捷性和准确性设计对比实验: 取乳制品类、酒类、茶类 3 种食品类别为采集对象, 由政府部门、企业、人民网三方各自独立设置关键词, 以获得的 3 份数据代表传统人工采集方法, 其中政府部门由贵州省分析测试院的工作人员为代表, 企业方由食品安全与营养 (贵州) 信息科技有限公司的工作人员为代表. 另外, 再运用网络爬虫技术, 使用 Python 的 requests 库解析页面数据接口获取相关数据, 采集新浪微博中乳制品类、酒类、茶类 3 种食品类别的信息, 以获得的数据代表线下流行的分析挖掘方法. 将得到的 4 组数据与监控探针采集到的数据做比照, 比较 5 组数据的挖掘时间、有效数据量、无效数据量等指标, 从而验证监控探针采集数据的速度和效率.

3.1 乳制品类

政府部门自设关键词“奶粉”“三聚氰胺”和“北京”; 企业自设关键词“酸奶”“乳酸菌”和“发酵”; 人民网自设关键词“奶茶”“肥胖”和“危害”; 运用基于贝叶斯网络模型的食品安全舆情监控探针算出所致食品安全风险概率偏高的 3 个关键词: “乳制品”P=95/23264×0.95=0.39%、“乳基婴儿配方食品”P=5/23264×0.95=0.02%、“奶酪”P=3/23264×0.95=0.01%, 因此设 3 个关键词为“乳制品”“乳基婴儿配方食品”和“奶酪”. 挖掘时间为 3 s, 较传统法人为设置关键词使用时间 (政府 15 s、企业 12 s、人民网 10 s、网络爬虫技术 9 s) 明

显缩短; 产生的垃圾数据仅 9 条, 较传统法人为设置关键词产生的垃圾数据 (政府 52 条、企业 512 条、人民网 159 条、网络爬虫技术 47 条) 明显减少; 数据有效率为 83.6%, 较传统法人为设置关键词 (政府 54.3%、企业 54.7%、人民网 56.2%、网络爬虫技术 63.0%) 准确率明显提高 (见表 3).

3.2 酒类

政府部门自设关键词“酒类”、“发酵”和“工艺”; 企业自设关键词“啤酒”“青岛”和“生产”; 人民网自设关键词“葡萄酒”“张裕”和“发酵”; 运用基于贝叶斯网络模型的食品安全舆情监控探针算出所致食品安全风险概率偏高的 3 个关键词: “酒类”P=2299/3264×0.95=9.39%、“白酒”P=25/23264×0.95=0.1%、“黄酒”P=7/23 264×0.95=0.03%, 因此设关键词为“酒类”“白酒”和“黄酒”. 挖掘时间为 2.5 s, 较传统法人为设置关键词使用时间 (政府 13 s、企业 14 s、人民网 5 s、网络爬虫技术 6 s) 明显缩短; 产生的垃圾数据仅 9 条, 较传统法人为设置关键词产生的垃圾数据 (政府 257条、企业 785 条、人民网 28 条、网络爬虫技术 69 条) 明显减少; 有效率为 77%, 较传统法人为设置关键词 (政府 55.9%、企业 52.3%、人民网 68.2%、网络爬虫技术 58.9%) 准确率明显提高 (见表 3).

3.3 茶类

政府部门自设关键词“茶类”“工艺”和“检测”; 企业自设关键词“绿茶”、“红茶”和“销售”; 人民网自设关键词“茶类”、“加工”和“贮存”; 运用基于贝叶斯网络模型的食品安全舆情监控探针算出所致食品安全风险概率偏高的 3 个关键词: “茶叶及其制品”P=7/23264×0.95=0.03%、“绿茶”P=29/23264×0.95=0.12%、“红茶”P=165/23264×0.95=0.67%, 因此设关键词为“茶叶及相关制品”、“绿茶”和“红茶”. 挖掘时间为 2.4 s, 较传统法人为设置关键词使用时间 (政府 15 s、企业 10 s、人民网 7 s、网络爬虫技术 11 s) 明显缩短; 产生的垃圾数据 64 条, 较传统法人为设置关键词产生的垃圾数据 (政府 29 条、企业 381 条、人民网 23 条、网络爬虫技术 45 条) 明显减少; 有效率为 77.9%, 较传统法人为设置关键词 (政府 52.5%、企业 55%、人民网 59.6%、网络爬虫技术 55.9%) 准确率明显提高 (见表 3).

表 3 食品安全舆情数据采集监控探针与传统方法对比实验

食品类别	检索方法	样本量 (条)	挖掘时间 (s)	有效数据量 (条)	垃圾数据量 (条)	有效率 (%)
乳制品类	政府部门自设关键词	114	15	62	52	54.3
	企业自设关键词	1130	12	618	512	54.7
	人民网自设关键词	363	10	204	159	56.2
	网络爬虫技术	127	9	80	47	63.0
	监控探针	55	3	46	9	83.6
酒类	政府部门自设关键词	583	13	326	257	55.9
	企业自设关键词	1644	14	859	785	52.3
	人民网自设关键词	88	5	60	28	68.2
	网络爬虫技术	167	6	98	69	58.9
	监控探针	39	2.5	30	9	77
茶类	政府部门自设关键词	61	15	32	29	52.5
	企业自设关键词	846	10	465	381	55.0
	人民网自设关键词	57	7	34	23	59.6
	网络爬虫技术	102	11	57	45	55.9
	监控探针	289	2.4	225	64	77.9

4 结论与展望

基于贝叶斯网络的食品安全舆情监控探针结合食品安全关键词库与贝叶斯网络概率算法, 运用贝叶斯网络模型推理食品安全风险概率大小, 定义节点并部署模型, 查询到“食物大类”“食品检测不合格”和“危害症状”导致的风险结果. 不仅评估出食品安全关键词库中的局部风险, 还可以根据示例中的计算与分析过程, 获取食品安全相应风险问题的全面评估, 实现高风险性关键词的优先采集, 有效解决了食品安全舆情监测数据采集中的不精准及效率低等问题.

本研究围绕发现问题、分析问题、解决问题的思路展开研究, 针对食品安全舆情监测系统数据采集环节所存在的问题提出科学假设. 首先利用 MySQL 数据库建立食品安全关键词库; 然后, 运用贝叶斯网络模型将关键词库构建形成监控探针, 并选择食品安全舆情监测系统进行数据采集; 最后, 以乳制品、酒及茶 3 种

食品案例的数据代入方法中与传统人工采集、网络爬虫技术形成对比实验, 通过对比数据挖掘时间和采集数据有效率, 验证假设成立. 由此得出, 关键词库引入贝叶斯网络模型形成监控探针, 可有效提高食品安全舆情数据采集的及时性与精准度, 精准定位不同的采集对象, 节约了舆情监测体系的采集成本, 拓展了食品安全网络舆情监测系统推广应用的范围.

参考文献

1 陶光灿, 刘学生, 夏虎, 等. 一种食品安全舆情监控方法及系统: 中国, 201811030986.1, 2019-01-18.

2 刘波维, 曾润喜. 我国食品安全网络舆情研究现状分析. 情报杂志, 2017, 36(6): 118–123, 166. [doi: 10.3969/j.issn.1002-1965.2017.06.021]

3 王旎, 孙晓红, 祁海峰, 等. 我国食品安全网络舆情监测系统应用难点及对策研究. 中国农业科技导报, 2021, 23(5): 8–17. [doi: 10.13304/j.nykjdb.2020.0475]

4 郭勋诚. 朴素贝叶斯分类算法应用研究. 通讯世界, 2019, 26(1): 241–242. [doi: 10.3969/j.issn.1006-4222.2019.01.157]

5 马欣鑫, 邓平科, 陈威屹, 等. 基于贝叶斯理论的多系统定位融合算法. 科学技术与工程, 2019, 19(26): 288–293. [doi: 10.3969/j.issn.1671-1815.2019.26.045]

6 肖绍武. 基于云计算的食品安全舆情分析算法研究 [硕士学位论文]. 贵阳: 贵州大学, 2018. 16–18.

7 滕越. 基于因果效应的贝叶斯网络结构学习方法及应用 [硕士学位论文]. 合肥: 合肥工业大学, 2019. 1–5.

8 王涛. 一种监控探针描述语言及其编译器的设计与实现 [硕士学位论文]. 长沙: 国防科学技术大学, 2009. 22–24.

9 于娟, 刘强. 主题网络爬虫研究综述. 计算机工程与科学, 2015, 37(2): 231–237. [doi: 10.3969/j.issn.1007-130X.2015.02.007]

10 耿贞伟, 保富. 网络环境下的大数据采集和处理. 软件工程, 2019, 22(6): 47–49.

11 霍福华. 关于大数据的数据处理探讨. 软件工程, 2019, 22(3): 32–34.

12 丁俊, 郑辉. 大数据时代下的动态可配置数据采集系统的研究与设计. 计算机应用与软件, 2018, 35(3): 75–79. [doi: 10.3969/j.issn.1000-386x.2018.03.014]

13 唐立. 基于文本挖掘的网络舆情监控与分析系统的研究与实现 [硕士学位论文]. 长沙: 湖南大学, 2016. 2–17.

14 丁祎姗, 杜彦辉, 朱衍丞, 等. 基于知识图谱的国内关键词抽取技术研究. 软件导刊, 2020, 19(2): 273–277.

15 袁明. 基于隐性主题模型和新词发现的关键词抽取研究 [硕士学位论文]. 北京: 北京邮电大学, 2014. 3–5.

16 潘晓英, 陈柳, 余慧敏, 等. 主题爬虫技术研究综述. 计算机应用研究, 2020, 37(4): 961–965, 972.

17 汤露阳. 面向网络舆情分析的数据采集与管理方法研究 [硕士学位论文]. 成都: 电子科技大学, 2017. 3–20.

18 Luhn HP. A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development, 1957, 1(4): 309–317. [doi: 10.1147/rd.14.0309]

19 赵京胜, 朱巧明, 周国栋, 等. 自动关键词抽取研究综述. 软件学报, 2017, 28(9): 2431–2449. [doi: 10.13328/j.cnki.jos.005301]

20 Witten IH, Paynter GW, Frank E, *et al*. KEA: Practical automatic keyphrase extraction. Proceedings of the 4th ACM Conference on Digital Libraries. Berkeley: ACM, 1999. 254–255.

21 张丽. 文本挖掘中关键词与文本摘要自动提取研究 [硕士学位论文]. 青岛: 青岛理工大学, 2018. 14–15.

22 Pearl J. Fusion, propagation, and structuring in belief networks. Artificial Intelligence, 1986, 29(3): 241–288. [doi: 10.1016/0004-3702(86)90072-X]

23 张渊. 基于改进的贝叶斯网络模型的齿轮箱故障诊断研究 [硕士学位论文]. 太原: 中北大学, 2019. 2–19.

24 丁华东, 许华虎, 段然, 等. 基于贝叶斯方法的网络安全态势感知模型. 计算机工程, 2020, 46(6): 130–135.

25 王森林. 基于监督式机器学习的疾病智能诊断算法研究与实现 [硕士学位论文]. 长沙: 湖南大学, 2019. 40–48.

26 梁书彤, 郭茂祖, 赵玲玲. 基于机器学习的医疗决策支持系统综述. 计算机工程与应用, 2019, 55(19): 1–11. [doi: 10.3778/j.issn.1002-8331.1903-0485]

27 王刚, 牛宏侠. 融合全局与局部特征的贝叶斯人脸识别方法. 计算机工程与应用, 2019, 55(11): 172–178. [doi: 10.3778/j.issn.1002-8331.1802-0037]

28 刘浩然, 丁攀, 郭长江, 等. 基于贝叶斯算法的中文垃圾邮件过滤系统研究. 通信学报, 2018, 39(12): 151–159.

29 陈珂馨. 基于贝叶斯算法的民航机票预测系统研究 [硕士学位论文]. 长沙: 湖南大学, 2017. 15–22.

30 北京人民在线网络科技有限公司. 人民众云. https://rmzy.peopleyun.cn/#/index/indexinit. [2021-03-01].

31 食品伙伴网. 食品标准. http://down.foodmate.net/standard/index.html. [2021-03-01].

32 食品安全与营养 (贵州) 信息科技有限公司. 食品安全云. http://community.fsnip.com/lims-standard-cloud/home/index.shtml. [2021-03-01].

33 国家食品药品监管总局. 《食品生产许可分类目录》. http://www.cnhfa.org.cn/fagui/show.php?itemid=11. (2016-01-22).

34 邓云, 王华. 供应链视角下食品安全风险因子分析. 江苏商论, 2019, (10): 3–9. [doi: 10.3969/j.issn.1009-0061.2019.10.001]

35 Tummalapalli S, Machavarapu VR. Managing MySQL cluster data using cloudera impala. Procedia Computer Science, 2016, 85: 463–474. [doi: 10.1016/j.procs.2016.05.193]

36 颜清, 苗壮, 赖鑫生, 等. 大数据时代关系数据库 MySQL 的创新与发展. 科技风, 2020, (20): 75–76.