****

**2020 浙江数据开放创新应用大赛**

食安浙江

一、需求分析

（一）用户需求分析

食品质量安全是“健康中国”和“质量强国”国家战略的重要组成部分，事关民生、民心，是各级政府的工作重心。本作品的目标用户为普通民众及终端消费者，拟解决的需求包括：1、普及食品安全知识，提高大众的科学素养和风险防范水平。2、促进监管信息透明公开化，提高民众的知情权和对食品事件的认知能力。

本作品聚焦食品检验中的有害化学物质，包括农药残留、兽药残留、霉菌毒素、食品添加剂及其他有害物质等。这些有害物质的检验是食品安全把控的关键环节。系统将为用户提供以下功能：1）科普百科。提供常见食品有害物质的化学信息、毒理信息、鉴别方法。2）知识图谱。通过关联规则学习和文本挖掘技术，获取各种食品和有害化学物的关联强度/支持度，形成食品实体和化学实体的知识图谱。通过交互式可视化技术，用户可以了解感兴趣食品易感的有害化学物质种类。3）食品安全事件的时空演化。通过实时挖掘食品安全事件，提取食品、化学物质、时间和空间等信息，实现时空可视化，便于用户掌握食品安全事件的区域性和季节性特征。4）全程溯源。通过与本省标杆企业的合作，全程采集并向用户公开食品检测数据，增加监管部门和生产企业的透明度。基于全过程检测数据构建GERT质量传导网络，实现风险发现和追溯功能。

（二）行业和竞品分析

目前，面向公众的食品安全应用以门户和主题网站为主，如“中国食品安全网”、各级政府和监管部门网站等。这些平台多采用Web 1.0的传统模式，主要作为新闻报道、政府公告等信息的公共发布平台。

从数据资源的角度看，与食品安全主题相关的各类数据目前以各种异构的、私有的、非结构化的形式分散在不同的网站和系统中，未能有效的集成和融合。这导致了1）不能发挥出大数据的价值，特别是缺少基于非结构化文本数据的智能化应用，如知识图谱、时空演化分析等。2）未集成监管部门和企业内部的检测数据，也未实现结构化及非结构化数据的深度融合，难以支持全过程溯源和全景式画像。

二、解决方案

（1）应用场景一：知识图谱

通过文本挖掘技术，从互联网各类文本中提取各种食品类型和相应化学物质的关联强度/支持度，形成食品实体和化学实体的知识图谱。进一步借助交互式可视化技术，用户可以直观了解特定食品和有害化学物质的关联强度/曝光频率。

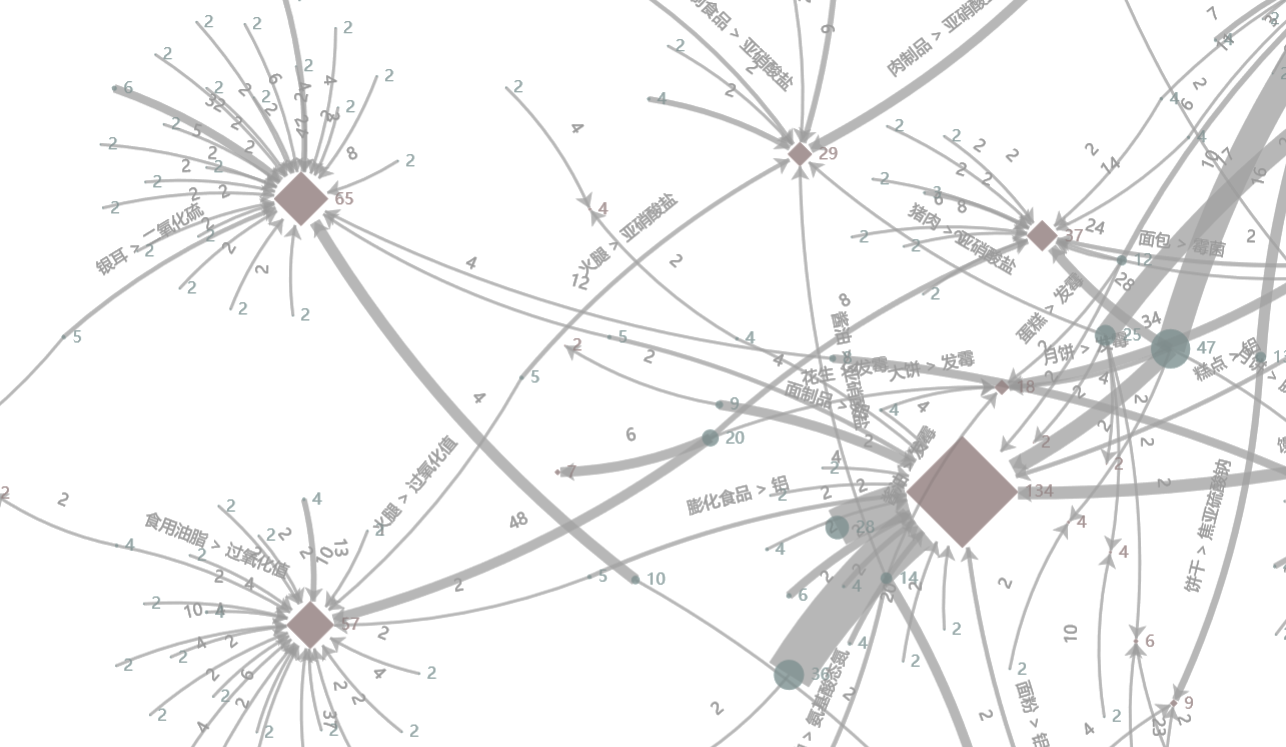


图 1 使用力导向图进行知识图谱的可交互可视化

应用场景二：时空演化可视化

从新闻报道、网络舆情等文本数据中实时提取食品安全事件的时间和地理信息，为特定的食品安全事件渲染时空演化过程。便于用户掌握重大事件的起源、发展和消亡，以及不同食品的区域性和季节性风险特征。

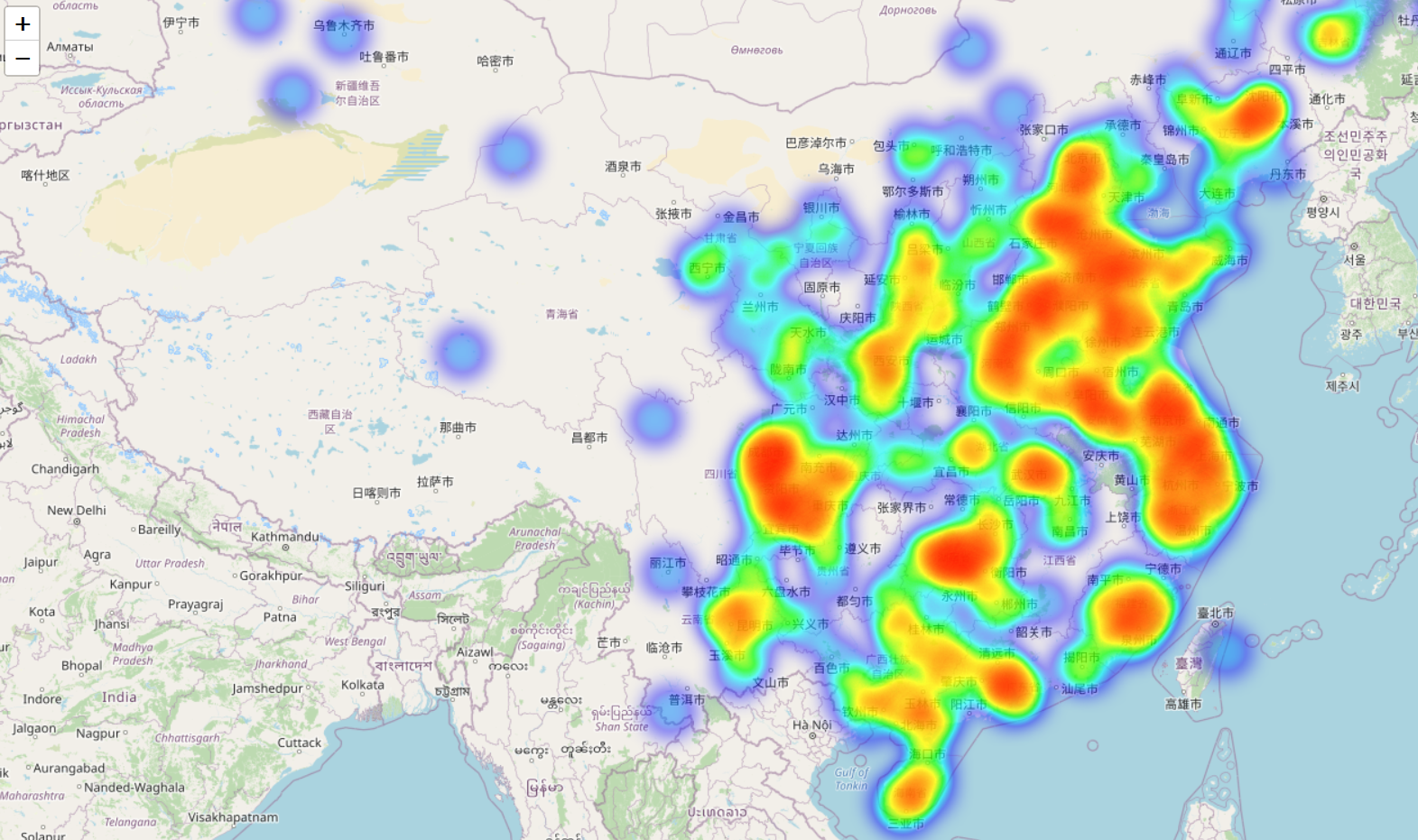


图 2 食品安全事件的空间分布

应用场景三：科普知识库

为用户提供常见的非法添加物百科知识，如化学信息、毒理信息、简易鉴别方法以及有关部门的抽检数据。

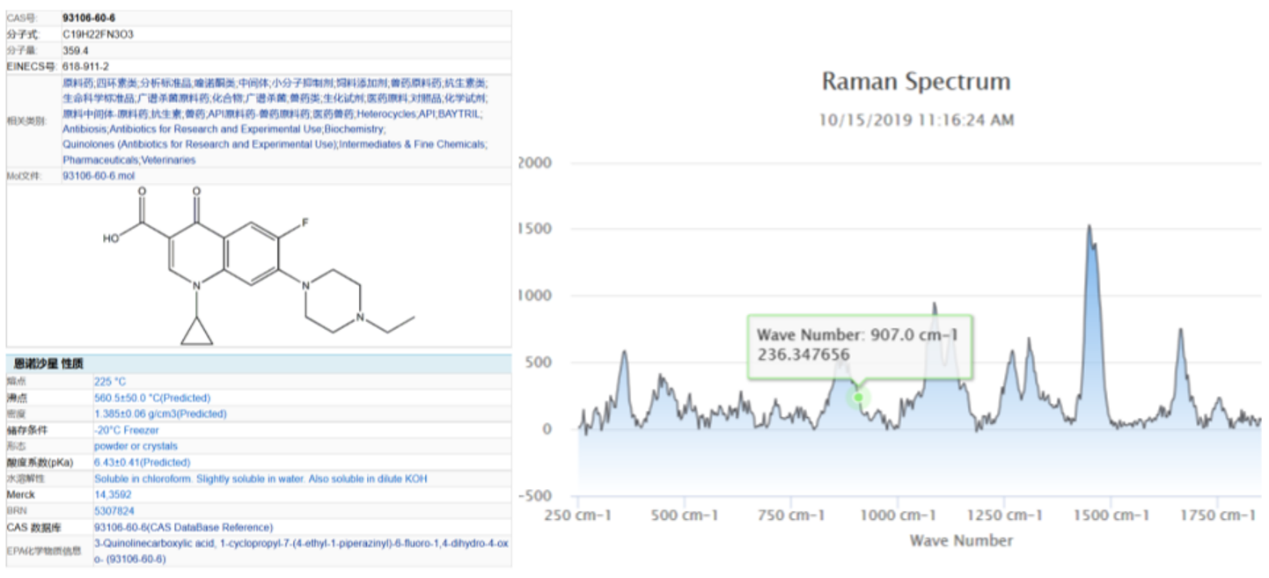


图 3 恩诺沙星科普页面和拉曼光谱数据

应用场景四：全过程溯源

通过与本省标杆企业的合作，采集并公开监管部门抽检数据和企业生产数据，提供包括食品有害物质检测在内的全过程溯源数据。同时，通过对检测数据和网络文本数据的融合分析，基于GERT（Graphic Evaluation and Review Technique，图形评审技术）技术构建食品全过程质量链、风险传播路径等质量模型，可以指导企业把控薄弱环节。

三、数据使用

（一）数据清单

数据一：浙江省政府公开数据集(data.zjzwfw.gov.cn)《食品监督抽查不合格信息》，文件名cata\_4370.xls，共1139条抽检数据。

使用字段：“不合格项目││检验结果││标准值”，字符串类型；“食品名称”，字符串类型。

数据作用：该文件提供了我省历年抽检不合格的食品和风险化学物质的名称，可用于构建初始的概念集和词典，服务于后续的命名实体提取和术语识别。

数据二：全国打击违法添加非食用物质和滥用食品添加剂专项整治领导小组颁布的五批《食品中可能违法添加的非食用物质和易滥用的食品添加剂名单》。共110条记录。

使用字段：食品名称、添加剂名称、鉴定方法

数据作用：食品名称和添加剂名称用于构建初始的概念集和词典，服务于后续的命名实体提取和术语识别。鉴定方法用于建设科普内容。

数据三：食品安全文本语料库，来源：中国食品安全网 – 曝光专栏、专家解读专栏；食安中国。共5599份文档。

使用字段：整篇文档。

数据作用：构建主题语料库；从文档中提取食品、添加剂、时间、空间等实体及相互之间的语义关联。根据所提取到的信息，不断更新知识库，用于后续的知识图谱等应用。

数据四：检测数据。来源：浙江工商大学浙江食品质量安全工程研究院、新希望集团、浙江盐业、大晟药业，及浙江省发展资产经营有限公司下属企业。

包括各类食品的检测数据，如电子鼻、飞行时间质谱仪、离子迁移谱仪、拉曼光谱仪、红外光谱仪、紫外光谱仪等；主要表征指标有：感官质量、理化质量、微生物质量等。目前积累了 6 万条谱图历史数据、超过 10 GB 数据量。

此数据集既作为检测的原始证据，也被用于训练和验证各类基于图谱数据的机器学习算法。

（二）算法模型与技术方案

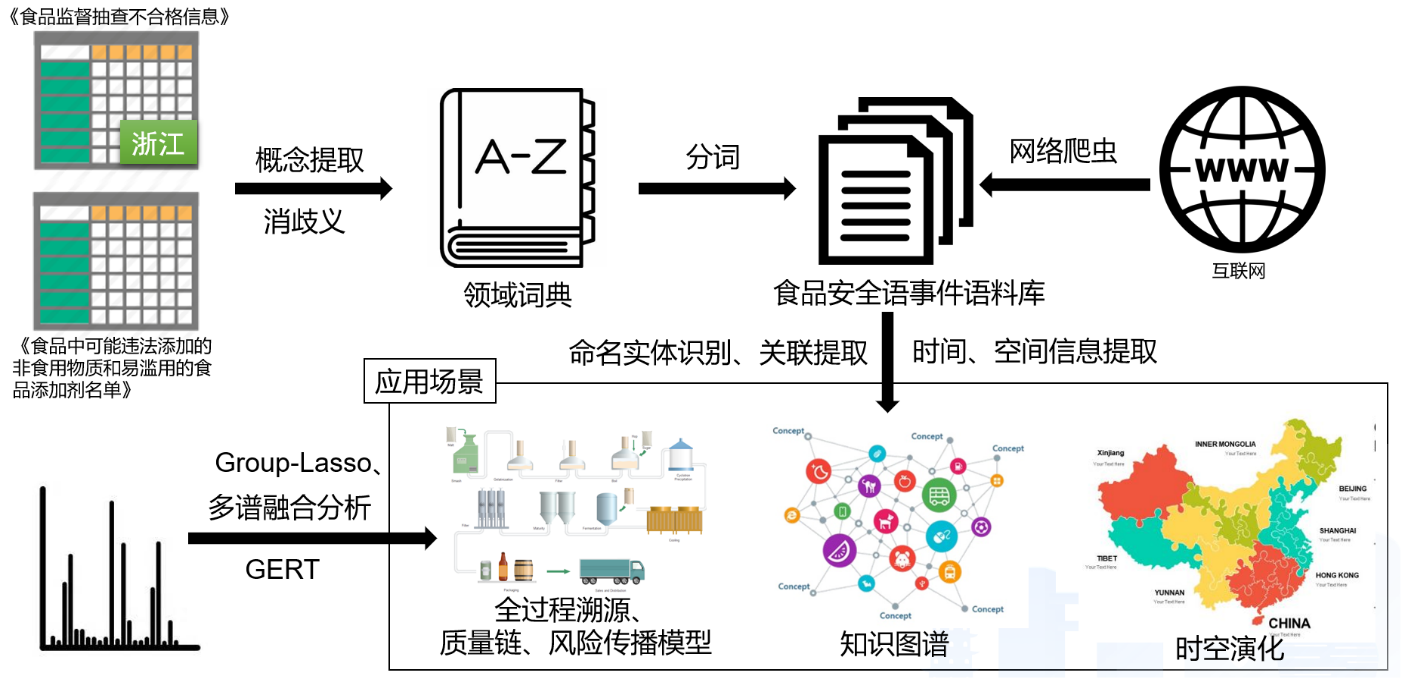


图 4 算法模型和整体技术方案

在本项目所使用的四种数据中，数据一、数据二、数据三属于非结构化数据。对此类数据的处理使用文本挖掘和自然语言处理技术。具体步骤如图 4所示，首先，从数据一和数据二整理出食品和有害化学物质的术语，构成基本的领域词典。然后，通过网络爬虫工具收集食品安全相关的新闻报道、事件通报、网购评论、消费者舆情等文本数据，构建食品安全事件主题语料库。进而，使用构建的领域词典配合通用的中文分词工具对语料库文本进行分词，从中提取食品、化学物质、时间、空间等实体及相互关系。最后，提取到的信息和关联更新到本地知识库，进一步用于构建知识图谱、时空演化分析等终端应用。

数据四属于结构化数据。此类数据具有独特鲜明的领域特点，特别是快检图谱类数据多表现出高维稀疏的特点，我们将重点研究基于化学先验知识的Group-LASSO特征选择算法以及多谱融合分析算法。此类结构化数据的分析结果作为食品各个阶段质量的量化评价指标，可以结合其它质量要素构建全过程质量链、风险传播等GERT模型。

（三）数据安全

可能涉及隐私的数据：网络评论、消费者舆情等网络文本。针对此类数据，在数据采集和数据清洗阶段将使用脱敏和匿名化技术处理。

四、作品价值

本作品的价值和意义在于：1、用户获得感：普及民众科学知识，提高科学素养和食品风险防范水平；2、治理效能：提升公共数据的透明度，提升政府和企业公信度；3、数字经济效益：充分发掘和利用海量文本大数据，通过构建全过程质量链和风险传播模型，为食品安全提供全景式大数据画像，改进企业质量管理；4、宏观战略：支持“健康中国”和“质量强国”国家战略。

食品安全已提升到国家重大公共安全层面。习近平指出，“确保食品安全是民生工程、民心工程，是各级党委政府义不容辞之责”。本作品契合了国家战略，也顺应了重大民生需求。本作品对于提高食品企业信誉和消费者信任，促进食品行业健康可持续发展，改善民生福祉，提升我省食品产业竞争力具有重要的理论与现实意义。