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。本作品服务于政府监管部门，重点聚焦食品检验中有害化学物质的风险管理。这些化学物质是食品监管的关键，既包括外源性的，如农药残留、抗生素残留、食品添加剂、重金属污染等，又包括内源性的，如因存储不当出现的霉菌毒素等。

对于政府监管部门而言，数据资源是准确分析和科学决策的基础，然而，目前与食品安全相关的数据资源存在以下问题：1、“数据孤岛”问题。各类数据以不同的私有格式分散在政府、企业和互联网中。2、数据融合和集成度低。难以形成全景式的“大数据画像”，导致监管部门无法获得整体性认知，不利于科学决策。3、缺少有价值的大数据应用，数据资源效益未能充分发挥。

针对以上问题，作品将实现一个面向食品安全风险管理的大数据融合和分析平台，并为终端用户提供以下功能：1）知识图谱。通过关联规则学习和文本挖掘技术，从政府抽检数据和新闻通报中，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，可以直观了解特定食品易感的有害化学物质种类。2）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，便于掌握食品安全事件的区域性和季节性特征。3）食品有害化学物质的快速检测。基于光谱、质谱数据，构建化学指纹特征知识库；设计高效的特征匹配算法，全面提高抽检通量和效率。4）全过程溯源。通过与本省标杆企业的合作，基于全过程检测数据构建 GERT 质量传导网络，实现风险发现和预警功能。

（二）行业和竞品分析

目前，食品安全方面的信息服务产品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，主要作为抽检报告、新闻报道、政府公告等信息的公共发布平台。

从前文所论述的背景问题看，这些竞品的核心问题主要是“数据孤岛”。与食品安全主题相关的各类数据以各种异构的、私有的、非结构化的形式分散在不同的网站和系统中，

未能深度融合，难以形成全景式的“大数据画像”，不能发挥出大数据的价值。

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。

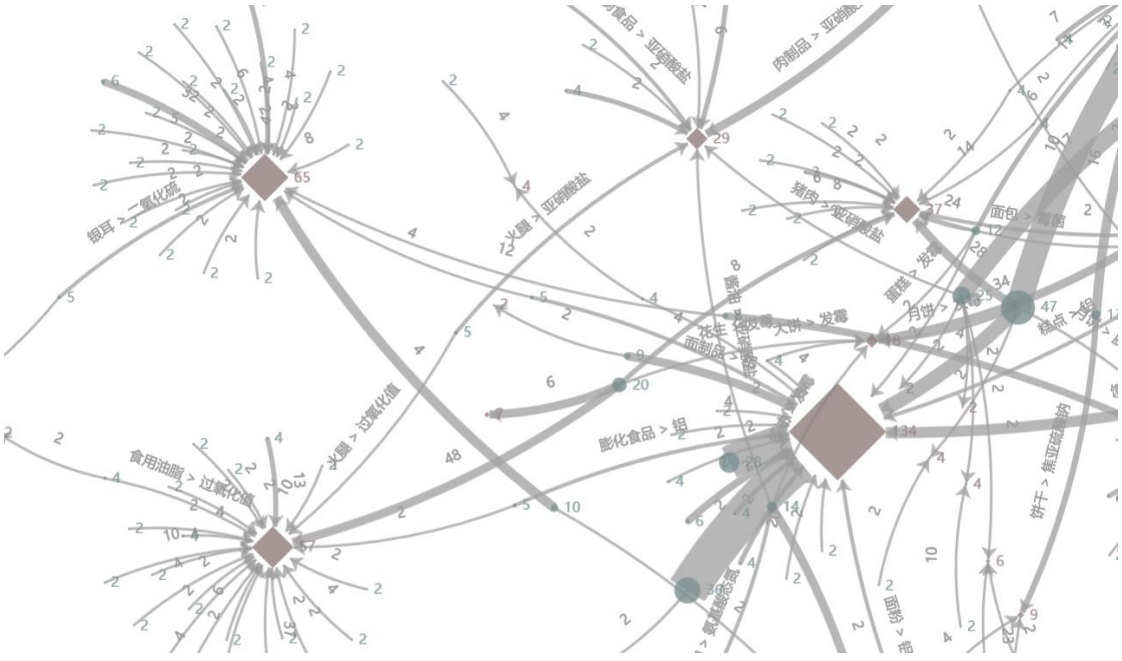

图 1 使用力导向图进行知识图谱的可交互可视化

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。

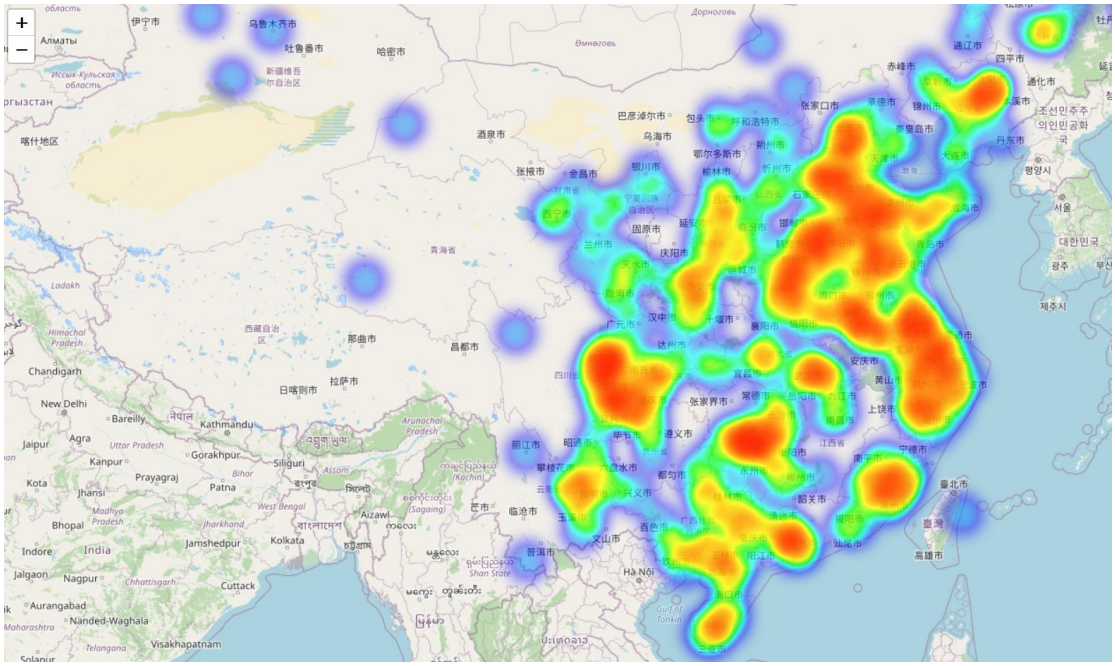

图 2 食品安全事件的空间分布

应用场景三：快速检测

基于团队前期积累的海量光谱、质谱等化学计量数据，构建有害化学物质的指纹特征图谱知识库。通过 Group Lasso 特征提取、多核 SVM 分类器、随机权神经网络等机器学习算法，训练适用于端智能场景的高效特征匹配和判别算法，实现终端场景的实时快检，以全面提高监管部门的检测效率。

应用场景四：全过程溯源

通过与本省标杆企业的合作，采集各个生产环节的质检数据，形成包括食品有害物质检测在内的全过程质量溯源数

据。同时，结合对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和风险传播路径。可以快速定位质量问题根源，协助企业和政府把控薄弱环节、提高风险管控水平。

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。

数据作用：作为食品质量检测的原始证据；构建化学指纹特征图谱知识库；用于训练适用于端智能场景的高效特征匹配和判别算法；结合生产过程中的其他质量要素，构建质量链和风险传播模型。

（二）算法模型与技术方案

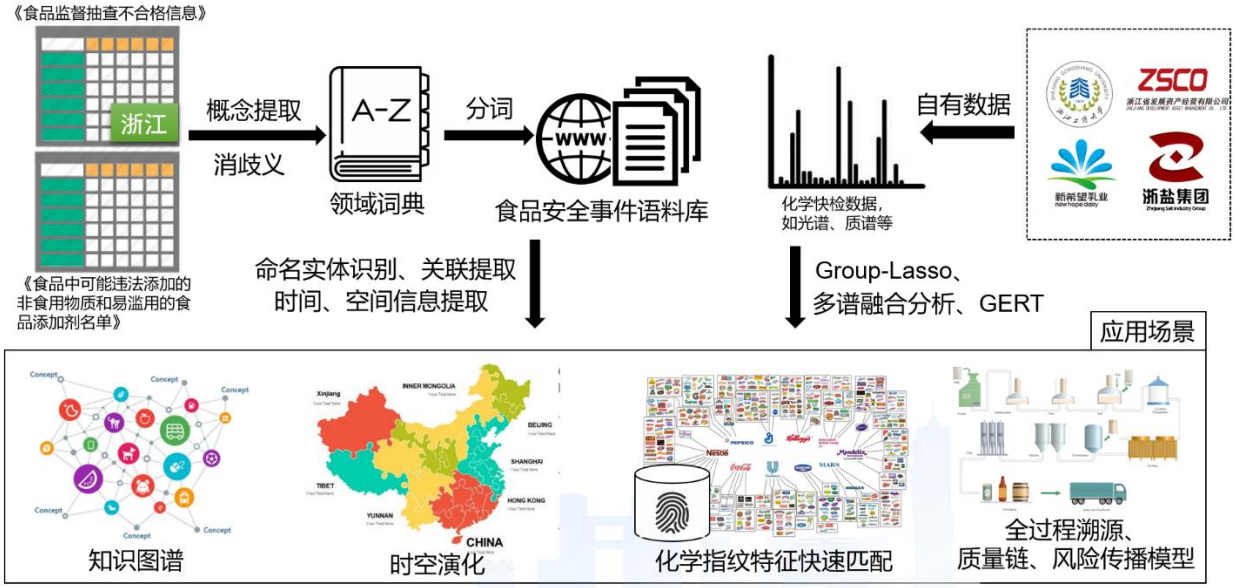

图 4 算法模型和整体技术方案

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、多核学习算法，以及多谱融合分析算法。此类结构化数据的分析结果用于构建化学物质的指纹特征图谱知识库，也可以用于计算食品各个阶段质量的量化评价指标，进一步结合其它质量要素构建全过程质量链和风险传播模型。

（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、用户需求：提供基于大数据的融合应用，提升食品安全监管效率。2、治理效能：打破“信息孤岛”，整合多源异构数据，形成协同的大数据资源。3、数字经济效益：大数据的价值发现，构建全过程质量链和风险传播模型，协助质量管理。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。本作品对于提高政府监管部门的执法效率和管理效能、

重建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，都具有重要的理论与现实意义。