大数据技术在疫情精准防控中的应用

李刚，郑佳，尹华山，黄文超
广州市数字政府运营中心，广东 广州 510623

摘要
以X市为例，针对超大城市的实际情况，基于大数据处理和分析方法，提出了基于“四标四实”数据建设疫情防控大数据库，并通过大数据技术辅助疫情防控的思路，建立了一套疫情态势实时感知、人员精准管控、企业精准帮扶的系统，对该系统中的数据建设状况和采用的关联规则挖掘算法、基于期望最大化概率聚类的感染预警机制和基于文本挖掘的非结构化数据利用策略等具体技术手段做了详细分析。该系统节约基层人力十余万小时、准确定位并跟踪到了重点人群上万人，为阻断疫情感染、提升企业复工复产率、减少经济损失起到了巨大作用，对各地通过大数据技术辅助疫情防控具有较大的借鉴意义。

关键词
疫情防控；大数据分析；四标四实

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2021009

Application of big data technology in precise prevention and control of epidemic situation

LI Gang, ZHENG Jia, YIN Huashan, HUANG Wenchao
Guangzhou Digital Government Operation Center, Guangzhou 510623, China

Abstract
Taking City X as an example, based on the actual situation of a mega-city, big data processing and analysis methods, a large database for epidemic situation prevention and control based on “four standards and four realities” data was built. And through big data technology, to assist epidemic situation prevention and control, a system of real-time awareness of the epidemic situation, precise personnel control, and precise enterprise assistance was built. The specific technical methods were analyzed in detail, such as the data construction status in the system, the association rule mining algorithm adopted, the infection warning mechanism based on expectation maximum probability clustering, and the unstructured data utilization strategy based on text mining. The system approximately saved more than 100 000 hours for country cadres, precisely located and traced tens of thousands of susceptible people who were focused, had played a huge role in blocking epidemic infection, elevating the rate of production resumption, and reducing economic losses, therefore it has a reference significance for all parts of the country.

Key words
prevention and control of epidemic situation, big data analysis, four standards and four realities

1 引言

2020年年初，由新型冠状病毒肺炎（COVID-19）带来的全球性疫情对我国各地造成了巨大冲击。在疫情全球流行的背景下，我国疫情得到了有效控制，这是我国政府治理能力整体提升的表现。自党的十八届三中全会提出“推进国家治理体系和治理能力现代化”以来，各地不断加强政务信息系统统筹和整合，强化数据资源的汇聚和分析利用，政务信息化不断朝数据化、智能化方向发展。

X市由于人口规模大、外来流动人口比例大、进出口繁荣，在疫情期间遭受到严重的冲击。但是，总体来说，X市实现了对疫情的有效防控，有序开展复工复产，保证了经济的平稳复苏，这很大程度上要归功于长久以来数字政府领域的积累和沉淀，比如以“四标四实”（即标准作业图、标准建筑物编码、标准地址库、标准基础网格，实有人口、实有房屋、实有单位、实有设施）为核心的基础信息采集和大数据库建设工作。

大数据技术应用为疫情数据的分析利用提供了重要的技术工具。通过对“四标四实”数据、重点人群数据、市民填报数据、基层摸查数据等不同来源的数据进行清洗、比对和挖掘分析，发现疫情线索，生成预警信息，为基层人员核查和辅助领导决策提供了重要的技术支撑。

下面针对疫情防控大数据建设及应用、大数据分析和挖掘技术在疫情防控中的应用两个方面，介绍基于大数据疫情防控的一系列行之有效的方法和技术。

2 疫情防控大数据建设及应用

2.1 “四标四实”基础数据建设情况

为推进“平安城市”建设，X市于2017年开始实施“四标四实”专项工作，建设了“数字政府”基础应用平台，制定了《四标四实专项信息共享目录》，并依托政务信息共享平台汇集了35个部门及11个区的数据，对全市道路、街巷名称不规范（包括无名、重名、一路多名、不标准）情况进行了全面排查清理，由民政部门依法确定的标准地名和公安机关依法确定的标准门楼牌组合生成标准地址。依托“标准作业图”，全面采集实有人口数据，实现人员、房屋、地址精准关联匹配，解决户籍人员存在的“一人多址、人户分离”、流动人口存在的居住登记和注销问题，为卫生、消防、公安、税务、交通、社保、城建、统计等各领域的政府服务提供了强有力的支撑。

X市通过“四标四实”工作汇聚了公安、住建、规划、国土、交通、民政、水务、环保、农业等35个职能部门的与自然人相关的信息，它不仅包含居民身份、房屋地址等基础信息，而且涵盖了人房居住关系，人口流动情况，常住人口工商登记、社保缴纳、就业、医疗，居民日常出行等个人全景式数据信息。目前，数字政府基础应用平台汇集超过2.5亿条城市基础数据，划分出近2万个城乡“标准基础网格”，定位视频点152万个，将全市人、房、业信息核准、更新后纳入“四标四实”大数据库。数字政府基础应用平台与市场监管等26个部门的应用系统进行对接，政府部门通过应用平台实行数据交换和更新，全面优化了基层治理能力，是全市治理能力现代化的里程碑性工程。

2.2 疫情防控大数据应用

在疫情期间，该市以数字政府基础应用平台和“四标四实”大数据库为依托，借助云计算、大数据技术，通过数据高度共享、系统高度融合、服务高度集成，建成疫情态势实时感知、人员精准管控、企业精准帮扶的疫情防控指挥系统。该系统支撑疫情监测分析、防控救治、资源调配，有力地支持疫情防控和复工复产政策措施快速部署、快速落地，逐步成为全市数据枢纽和决策指挥“智慧大脑”。疫情防控指挥系统进一步整合汇聚15个部门的22类数据，建立畅通的数据通道，持续将确诊人员、重点人员、集中观察点等疫情防控相关数据与“四标四实”数据进行全面关联，实现防控对象、防控设施精准上图，形成疫情指挥“一张图”。目前“一张图”已汇聚各类信息2.76亿条，通过小程序上报信息4 000多万条（含线索5万多条），监控重点人群（包括患者、密切接触者、集中观察人群）相关数据超过30万条。

基于“四标四实”的精准疫情防控模块，以“四标四实”大数据库的数据为基础，进一步汇聚整理了人房居住关系数据、政企事业单位数据以及单位从业人员数据、社保缴纳数据，建立了人员群居关系、人口家庭关系、工作同事关系等数据单元。疫情期间再次叠加确诊人群数据和红码人群数据、公共交通乘坐记录等，建立数据规则模型，精准识别重点人群，辅助防疫人员进行重点跟踪和布防。疫情防控大数据建设及应用情况图1所示。

平台根据业务数据类型，建立家庭人群、同住人群、同事人群、同楼人群、同社区人群、同行人群等数据实体，以确诊人员、疑似病例为核心，以发现时间和隔离要求为辅助条件，设置相关算法规则，精准识别高危人员、重点人员以及应跟踪观察的人员等不同级别的群体。

3 大数据分析和挖掘技术在疫情防控中的应用

为了充分利用数据中的隐含信息，有效

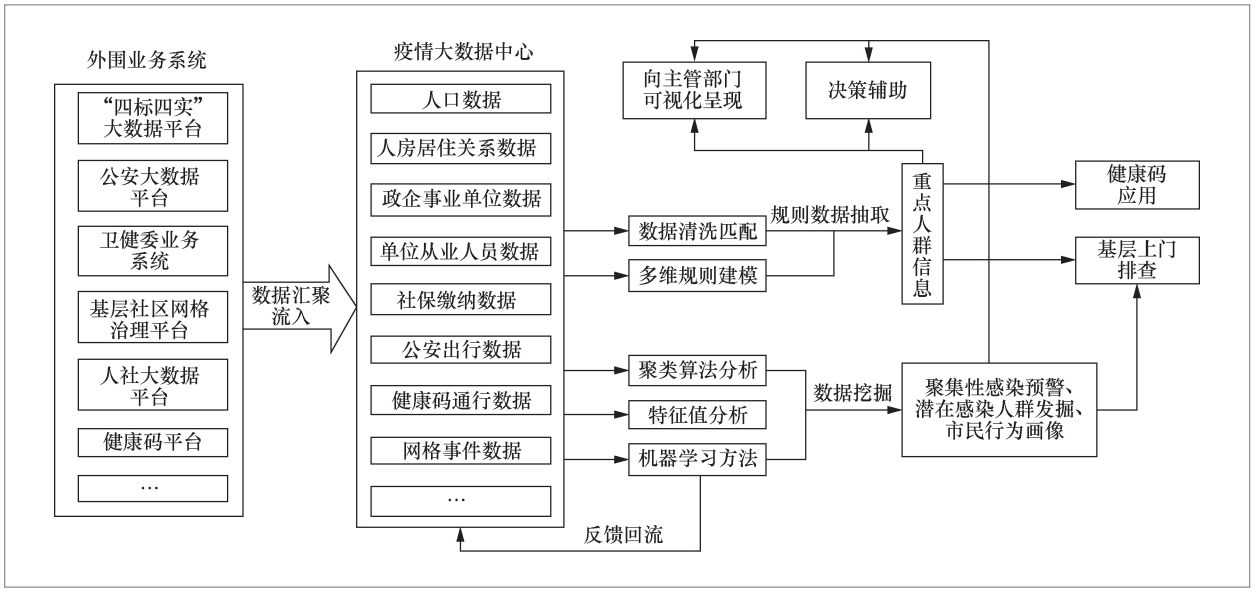

图1 疫情防控大数据建设及应用情况

识别潜在的高危人群、感染人群，对聚集性感染事件进行预警，必须采用关联规则、聚类分析等大数据分析和挖掘技术。X市疫情防控指挥系统构建了一套处理数据、挖掘数据的解决方案，并且在疫情防控实战中通过不断训练增强了自身的应对能力。

3.1 基于关联规则的重点怀疑对象挖掘

在“一张图”中，人与人的关系分为同住关系、同事关系、同乘交通工具等，而现实情况更为复杂，大多数关联关系没有被人为编写的关系数据库所收纳。然而这些关系造成的接触正是疫情防控工作中的盲点、难点。如何基于已有可接触的数据，推知间接的、隐含的、可以造成人员之间接触的关联关系，是应用数据挖掘的重点和难点。

应用关联规则发现目标数据的经典案例出现在零售领域，即耳熟能详的“啤酒与纸尿裤法则”，尽管在逻辑上难以推测出这两种商品的消费关联性，但是可以通过统计数据的积累，基于贝叶斯概率得到量化的二者间的关联度[1-2]。COVID-19感染人员传播途径时空分析**图2**所示。基于COVID-19感染人员传播途径时空分析的关联规则聚类如**图3**所示。

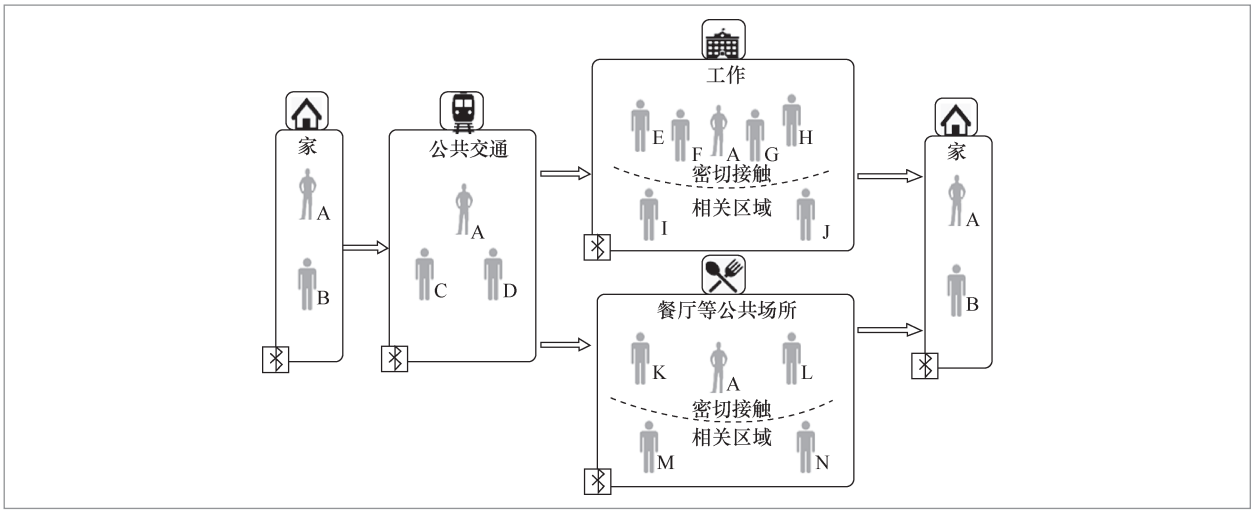

图 2 COVID-19 感染人员传播途径时空分析

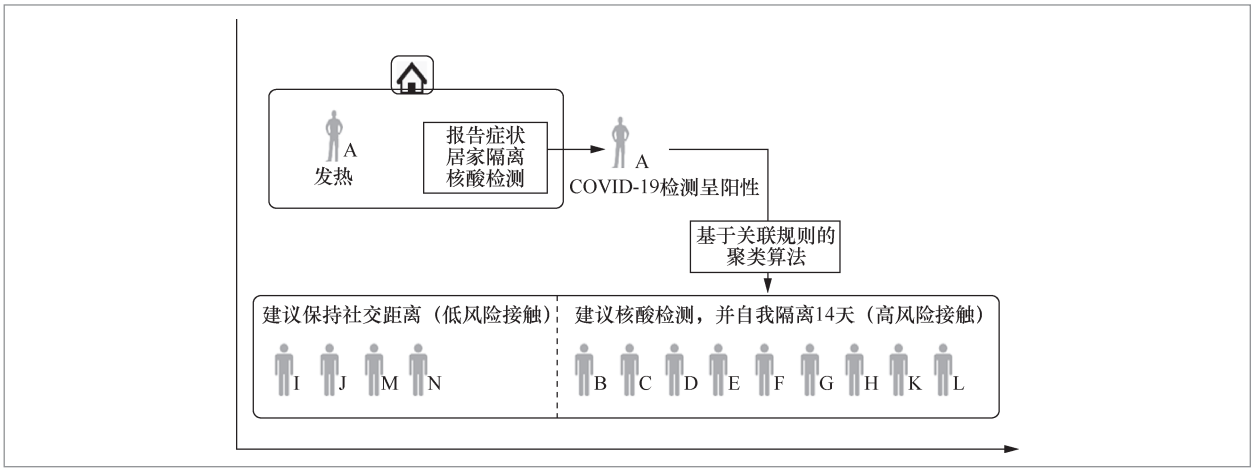

图 3 基于 COVID-19 感染人员传播途径时空分析的关联规则聚类

在疫情防控中，经常有这样一种情况：数据库中有一位市民的居住地和工作地数据，也有其工作时间（如进入单位时健康码扫码所得）数据，没有该市民的出行行程信息，但是该信息也是非常受关注的，关联规则挖掘就可以用在此处[3-4]。通过对具有隐含信息的情报进行分析，获取与其关联的信息，比如，通过工作地、居住地信息，结合规则的抽取算法，可以推知较可能的出行方式和涉及的公共交通线路等信息，从而在这一群体中出现感染者或疑似感染者时，能够快速对这一群体进行预警和监控。这一过程的意义在于为后续的重点人群筛选算法提供人群间关系的支持度和置信度数据。比如，某市民的居住地和工作地都在某地铁线路附近，并且工作时间已知。通过对地理信息的分析，模型可以针对其通过地铁出行，以及在某一时段在某一地铁线路上出现的概率给出判断，从而在该出行群体出现敏感对象时，量化该市民和该对象间的关系。

基于关联规则挖掘疫情重点人群的模型充分利用了已有的数字政府基础应用平台的数据资源，训练获取了包括家庭住址同居人、工作单位同事、通勤可能接触人群、常去消费场所可能接触人群、居住地附近活动可能接触人群等一系列关联关系数据，如**图4**所示。基于这种拓展，有效扩大了原本简单的关联规则，扩大了疫情紧急情况发生时的监控范围，有效地防止了疫情扩散。

为了科学地制定市民关联数据规则，本文采用基于规则模型的阈值数据抽取方法。对于每一个市民，在“四标四实”的基础数据库中，通过主成分分析、因子分析及基于机器学习的回归算法等特征提取方法，建立与防疫相关的市民关联向量，记为 $\boldsymbol{f}=(f_1,f_2,\cdots,f_m)$，其中每个维度对应家庭住址、工作单位、通勤方式、常去消费场所、居民活动时间、重要人流密集区域、隐藏同居人和密接同行人等一系列关联关系。假设市民A为疑似感染者，则可以定义 $\boldsymbol{f}_A=(f_{1\mathrm{A}},f_{2\mathrm{A}},\cdots,f_{m\mathrm{A}})=(1,1,\cdots,1)$。基于样本性质，可以定义市民B的关联向量为：如果A与B的第i个分量属于同一个范畴（可根据“四标四实”中的距离标定数据给出），则B继承A在该分量的值；否则，对应分量值满足 $\boldsymbol{f}_{j\mathrm{A}}=2\boldsymbol{f}_{j\mathrm{B}}$，即：

$$\boldsymbol{f}_{\mathrm{B}}=(f_{1\mathrm{B}},f_{2\mathrm{B}},\cdots,f_{m\mathrm{B}})=\left(\left(\frac{1}{2}\right)^{l_1}f_{1\mathrm{A}},\left(\frac{1}{2}\right)^{l_2}f_{2\mathrm{A}},\cdots,\left(\frac{1}{2}\right)^{l_m}f_{m\mathrm{A}}\right) \quad (1)$$

其中，$l_i=\begin{cases}1,\|\boldsymbol{x}_{i\mathrm{A}}-\boldsymbol{x}_{i\mathrm{B}}\|<\varepsilon\\0,\|\boldsymbol{x}_{i\mathrm{A}}-\boldsymbol{x}_{i\mathrm{B}}\|\geqslant\varepsilon\end{cases}$，$\boldsymbol{x}_{i\mathrm{A}}$、$\boldsymbol{x}_{i\mathrm{B}}$ 分别为“四标四实”中市民A、市民B的地理位置向量，ε 为阈值，$\left(\frac{1}{2}\right)^{l_i}$表示对应分量值的取值，由是否超过阈值决定。由上述递推关系，可以依次给出与市民A相关的关联信息网（correlation information network，CorNet）。在上述关联信息网中，每一个固定人口和流动人口都是网络中的一个节点，节点与节点之间的相关性可以通过高维高斯模型函数进行确定。

针对疫情人员关联信息网中重点人群的抽取过程，依据疫情发展的不同阶段设计出两种对应的数据抽取方法。

（1）阈值信息法

在疫情发展平缓的情形下，COVID-19检测呈阳性人员较少，病毒传播人群结构信息较明确，市民关联信息较易获取，故可以采取阈值信息法筛选重点怀疑对象人群。假设市民A为检测呈阳性人员（信息由X市相关医疗机构提供）。

第一步：设置地理位置阈值信息 ε_G，根据欧几里得范数多维球面区域 $R_{\mathrm{First}}=\{\mathrm{I}\,|\,\|\boldsymbol{x}_{o,\mathrm{A}}-\boldsymbol{x}_{o,1}\|_E<\varepsilon_G\}$ 初步筛选怀疑

对象人群，其中$\boldsymbol{x}_{\mathrm{o,A}}$、$\boldsymbol{x}_{\mathrm{o,I}}$分别为市民A和筛选对象I的原始个人地理位置信息，二者信息间距取欧几里得范数E。

第二步：求解市民关联模型阈值信息ε_F，根据非线性高斯不规则球面区域$R_{\mathrm{Final}}=\{\mathrm{J}\,|\,\mathrm{Cov(CovNet)}_{\mathrm{AJ}}<\varepsilon_F\}$进行确定，$\mathrm{Cov(CovNet)}_{\mathrm{AJ}}$表示筛选对象J与市民A在关联信息网中的节点相关性。市民关联模型阈值信息ε_F由传染病微分方程模型——易感者-感染者-易感者（susceptible-infectious-susceptible，SIS）模型确定：设$S(t)$为t时刻的易感者人数，$I(t)$为t时刻的感染者人数，N为群体总人数，则SIS模型可以表示为：

$$\begin{cases}\dfrac{\mathrm{d}S}{\mathrm{d}t}=-r\beta S\dfrac{I}{N}+\gamma I\\[2ex]\dfrac{\mathrm{d}I}{\mathrm{d}t}=r\beta S\dfrac{I}{N}-\gamma I\\[2ex]S(0)=S_0\\I(0)=I_0\end{cases}\qquad(2)$$

其中，r表示在单位时间内感染者接触到的易感者人数，β表示传染率，γ表示康复率，则市民关联模型阈值信息ε_F可以通过式（2）的解表示为：

$$\varepsilon_F=\lim_{t\to\infty}S(t)=\lim_{t\to\infty}(N-I(t))=\frac{N\gamma}{r\beta}\qquad(3)$$

（2）奇异值分解法

在疫情刚刚出现及新的突发情况出现的情形下，COVID-19感染人员情况不明，病毒传播人群结构信息较为模糊，市民关联信息很难获取，需筛查的重点人群目标不明确，因此利用奇异值分解法（singular value decomposition，SVD）筛选重点怀疑对象人群。假设市民A为检测呈阳性人员（信息由X市相关医疗机构提供）。

第一步：设置地理位置阈值信息ε_G，根据欧几里得范数多维球面区域$\tilde{R}_{\mathrm{First}}=\{\mathrm{I}\,|\,\|\boldsymbol{x}_{o,\mathrm{A}}-\boldsymbol{x}_{\mathrm{o,I}}\|_E<n\varepsilon_G\}$初步筛选怀疑对象人群，其中$n$为大规模筛选怀疑对象人群模型的参数，$\boldsymbol{x}_{\mathrm{o,A}}$、$\boldsymbol{x}_{\mathrm{o,I}}$分别为市民A和筛选对象I的原始个人地理位置信息。

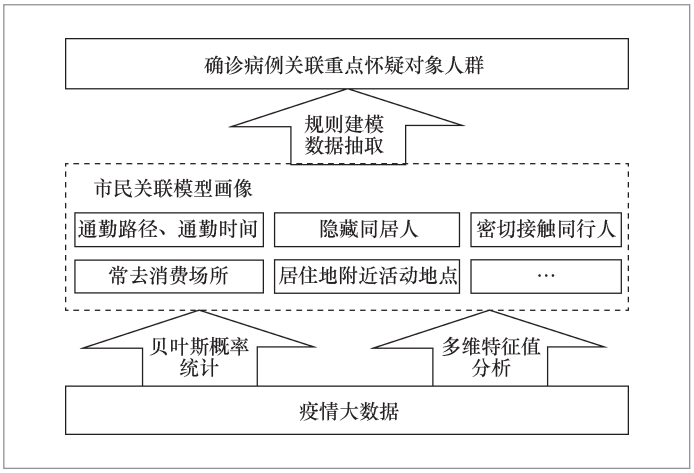

图4 基于关联规则的重点怀疑对象挖掘

第二步：根据阈值信息法中给出的市民关联模型阈值信息ε_F，将重点怀疑对象人群集合扩大为符合$\tilde{R}_{\mathrm{Final}}=\{\mathrm{J}\,|\,\mathrm{Cov(CovNet)}_{\mathrm{AJ}}<n\varepsilon_F\}$的非线性高斯不规则球面区域。针对$\tilde{R}_{\mathrm{Final}}$中的所有筛选对象J计算关联信息网的节点相关性$\mathrm{Cov(CovNet)}_{\mathrm{AJ}}$，构造大规模市民关联信息矩阵。由关联信息网节点相关性$\mathrm{Cov(CovNet)}_{\mathrm{AJ}}$的定义可知，$\boldsymbol{\Sigma}_{\mathrm{Bigdataselected}}$是一个对角线全为零的非负矩阵，因此由奇异值分解定理可知，存在矩阵$\boldsymbol{U}$、$\boldsymbol{V}$，满足：$\boldsymbol{\Sigma}_{\mathrm{Bigdataselected}}=\boldsymbol{U}^{\top}\boldsymbol{\Sigma}_{\mathrm{B}}\boldsymbol{V}$，其中$\boldsymbol{\Sigma}_{\mathrm{B}}=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_m,0,\cdots,0)$，并且$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m\geqslant0$。根据奇异值分解定理，矩阵$\boldsymbol{V}$的前$m_1$列恰为$\boldsymbol{\Sigma}_{\mathrm{Bigdataselected}}$的右奇异向量组，可以表示为$\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_{m_1}$，其中$m_1=\min\left\{m_k\in m\,|\,\dfrac{\sum_{i=1}^{m_k}\lambda_i\log_2\lambda_i}{\sum_{i=1}^{m}\lambda_i\log_2\lambda_i}>1-\alpha\right\}$为向量组中的向量个数，$\alpha$常取0.05，$m_1$表示右奇异向量组的子集中特值最大且加和几乎等于$\boldsymbol{\Sigma}_{\mathrm{B}}$之迹的最小向量组（SVD在此处起到筛选要点的作用，只考察最相关的奇异值，舍弃噪声）。因此大规模筛选重点怀疑对象人群的数据抽取方向确定为$\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_{m_1}$，向量组中每一个向量均表示一个与COVID-19感染

者密切相关的人员（即m_1个对象），$\boldsymbol{v}_i$表示其在关联信息网中各维度信息与市民A的距离向量。

3.2 基于概率聚类算法的聚集性感染预警

根据人工采集统计的疫情信息，仅可以发现家庭聚集、企业聚集等少数特征明显的聚集性感染情形，无法发现所采集条目文本外的隐含信息所关联的传播场景。例如当感染者在居民区附近的超市被感染时，这一信息不会被直接收入数据库中，但是可能可以用已有信息向量来表示。基于概率聚类算法，在高维度、大数据量的居民综合信息中，系统可以无监督地动态地发现具有高相似性的居民群体[5-6]，如果所得聚类在紧密性、信息距离等指标上符合要求，就会推送给人工审阅，如果的确是缺失的观察角度，就会被标注，进一步分析。通过这种方式，可以极大地扩充视野、查漏补缺，避免人工设计的不足[7-8]。

基于，期望最大化（expectation maxmum，EM）概率聚类的聚集性感染预警算法框架如图5所示。具体算法步骤如下。

第一步：根据“四标四实”基础数据平台的相关个人信息（主要包括地理位置信息（居住位置和工作位置为主）、网格普查所得行程信息、基于历史记录数据的行为画像等）建立个人行为向量。

第二步：将个人行为向量作为空间中的节点，建立基于EM概率聚类算法的聚集性感染预警模型。记个人行为向量的维数为d，设聚集性感染的情形有k种，其中包括家庭聚集、企业聚集以及商业场所聚集等情形。设第j个聚类集合χ_j的人行为向量集合可以表示为 $\boldsymbol{X}=\bigcup_{i=1}^{k}\chi_j$。假设基于$\theta=\{\boldsymbol{\mu},\boldsymbol{\Sigma}\}$的概率参数模型可以用来描述人行为向量集合的分布，其中θ为隐含参数集合，$\boldsymbol{\mu}$为参数各成分均值向量，$\boldsymbol{\Sigma}$为参数各成分协方差矩阵，则它的混合密度为：

$$p(h_k)=\sum_{j=1}^{c}p(h_k\mid\omega_j,\theta_j)P(\omega_j) \qquad (4)$$

其中，$p(h_k\mid\omega_j,\theta_j)$为条件概率，对应样本无偏的高斯参数估计为$\theta=\{\theta_j\mid j=1,\cdots,k\}$，$h_k\in\chi_j$，$\omega_j$表示混合模型中各成分的比例系数，$P(\omega_j)$为第$j$个聚类的先验分布函数，$c$为成分数目。上述混合概率分布对应的对数似然函数为：

$$L=\log[p(\boldsymbol{X}\mid\theta)]=\log\prod_{k=1}^{n}p(h_k\mid\theta)=\sum_{k=1}^{n}\log[p(h_k\mid\theta)] \qquad (5)$$

对上述对数似然函数对$\theta=\{\theta_j\mid j=1,\cdots,k\}$求偏导数可得：

$$\nabla_{\theta_i}L=\sum_{k=1}^{n}\frac{1}{p(h_k\mid\theta)}\nabla_{\theta_i}\left[\sum_{j=1}^{c}p(h_k\mid\omega_j,\theta_j)P(\omega_j)\right] \qquad (6)$$

根据$\{\ln[f(x)]\}'_x=f'(x)/f(x)$和$p(\omega_i\mid h_k,\theta_i)=p(h_k\mid\omega_i,\theta_i)P(\omega_i)/p(h_k\mid\theta_i)$，可以将上述方程化简为：

$$\nabla_{\theta_i}L=\sum_{k=1}^{n}p(\omega_i\mid h_k,\theta_i)\nabla_{\theta_i}\{\log[p(h_k\mid\omega_j,\theta_j)]\} \qquad (7)$$

根据大数定理，当样本量足够大时，样本集近似服从高斯分布。此时概率聚类模型参数集合$\theta=\{\theta_j=(\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)\mid j=1,\cdots,k\}$可以通过EM算法给出。

3.3 基于非结构化数据的文本挖掘发现疫情线索

疫情期间，来自政府服务热线、微信投诉平台、各部门投诉渠道和网站的疫情相关线索数量巨大、文本众多，各部门人力不足，无法通过人工充分利用这些非结构化的情报，使用文本挖掘的手段筛选文本数据中的关键信息、高频热词、舆情趋势以及把握群众心理十分有必要。

文本数据中包含的有价值的信息之一是与疫情有关的地理空间信息，结合疫情防控指挥系统整合的数据，系统可以快速定位疑似聚集性感染发生地、划定高危地区。利用BiLSTM+CRF模型进行文本序列标注，使用ERNIE语义模型进行实体抽取的训练和微调，所获得的模型可以识别出多级地址、主语、组织机构名和事项，从而从非结构化的文本中获得有价值的信息。为了增加模型的精确度和提高对本地情报信息的敏感度，在MSRA-NER数据集的基础上，训练分两步，即分别在源域进行学习和在目标域进行迁移学习，结合在先前积累的政务数据中筛选出的常见实体、事项、专有名词库，通过迁移学习，较快地得到了更精准的模型。

除地理信息等预定的需求实体信息外，十分常见而难以预先规定的重要信息是关于主体行为的谓词信息。对于少部分契合政务数据库中既有常见事项的信息，可以通过上述实体抽取技术获取，而由于疫情的突然性和新颖性，疫情线索中涉及的绝大多数情报并非既有事项，市民使用的口语化表述往往不利于统计和进行进一步的数据治理。因此，对日常语言化的表述进行“序列到序列（seq2seq）”的文本生成十分有必要。这种方法本质上和机械翻译使用的序列到序列的技术是类似的。针对汉语较复杂的短语、词组结构（比如由于断句的不同，句子可能产生歧义，这对文本生成是一个挑战），此处采用多层次注意力（multi-flow attention）机制的结构，即在词与词（word-by-word）和段与段（span-by-span）两个级别上的填充生成机制。于是，基于ERNIE-GEN模型[9]的结构，形成了一套文本缩写工具，使得系统可以将口语化的疫情线索文本简化为包含关键元素的简单陈述句，便于进一步的数据治理和人工汇总使用。

通过上述技术，计算机可以将一段文本翻译为只包含重点关注要素的简单陈述句或结构化的记录表，继而可以将之加入“一张图”的数据库中，或通过聚类等方法加以利用，如图6所示。

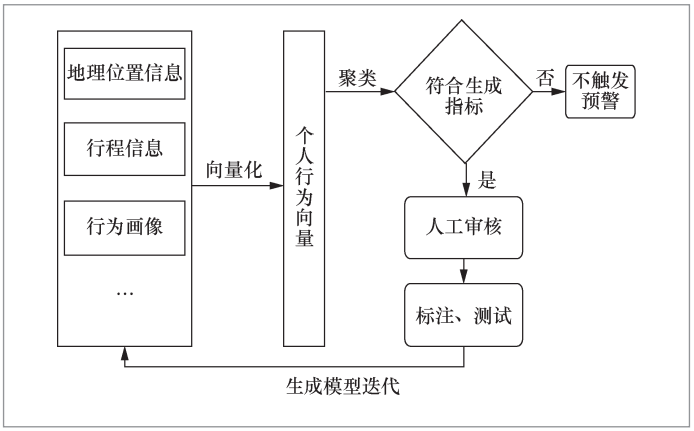

图 5 基于 EM 概率聚类算法的聚集性感染预警算法框架

在具体处理过程中，模型通过主题模型聚类方法，以及潜在语义分析、潜在狄利克莱分配（latent Dirichlet allocation）和概率潜在语义分析等手段发现等价词与主题的表示集合[7]。这种等价是通过文本内词之间的共现关系来实现的，特别适用于疫情文本非结构化数据的多主题标记特点[10]。而基于划分的聚类方法（包括K-均值、非线性K-均值和核K-均值等算法）可以通过角距离的度量实现非结构化大规模语义分类[11]。对于常见主题结构的文本语义分析，也可以通过词转向量的方法将非结构数据结构化，然后通过结构数据分析方法提取疫情线索信息。基于上述方法的算法结构如图7所示。

4 结束语

X市依托数字政府基础应用平台及“四标四实”基层治理数据，采用大数据分析和挖掘技术，快速构建了疫情防控指挥系统，在疫情防控工作中该系统发挥了

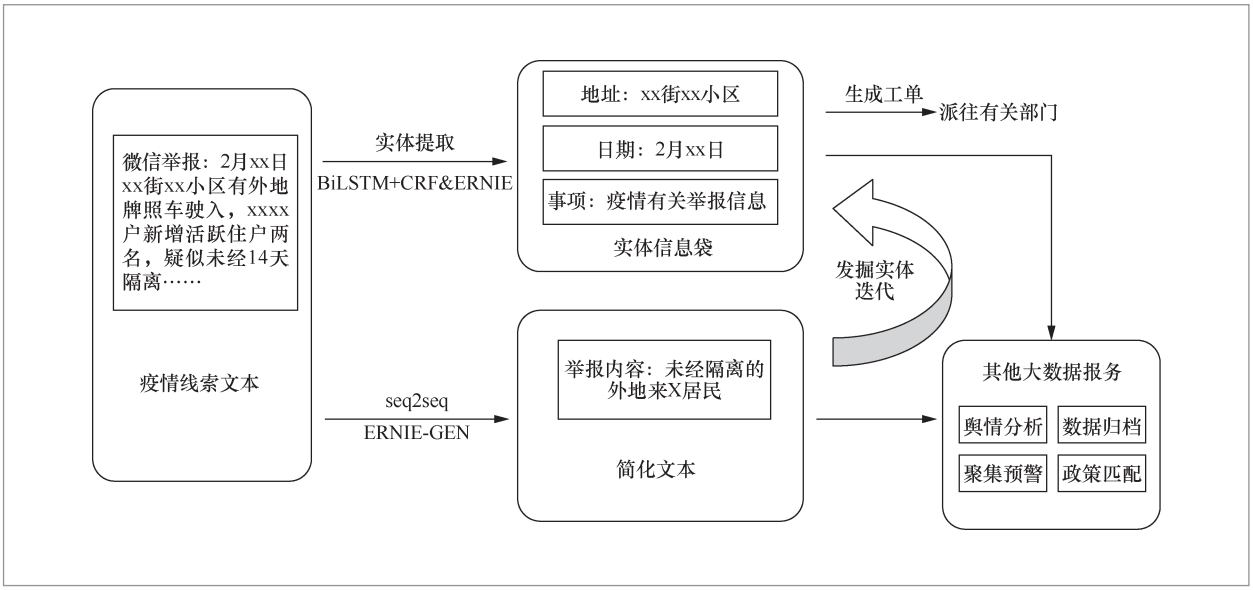

图 6 基于非结构化数据的文本挖掘发现疫情线索

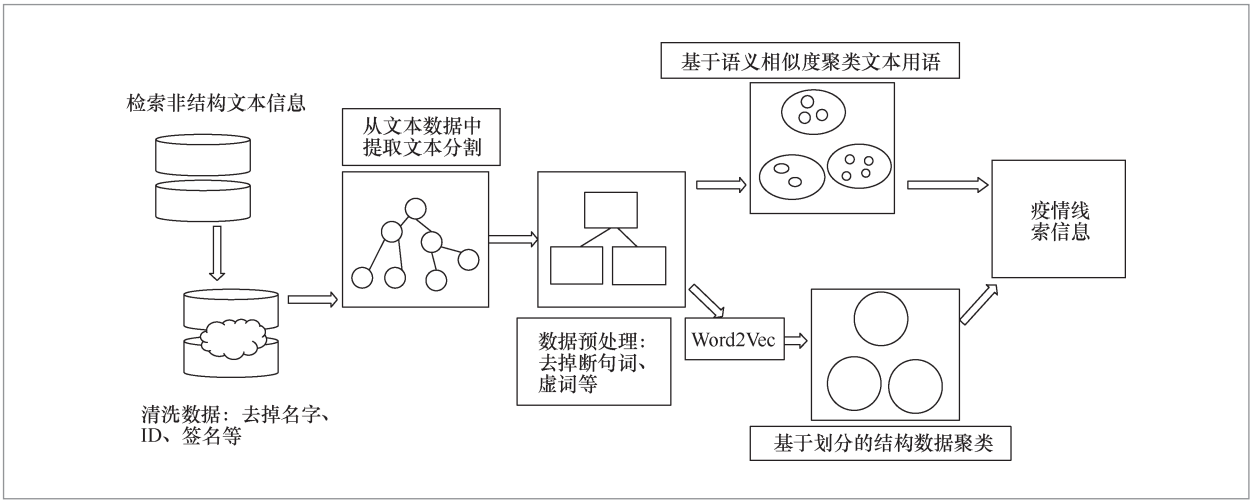

图 7 基于聚类方法的非结构化文本数据挖掘发现疫情线索算法结构

重要作用。该平台已应用于市委办公厅、市政府办公厅、市发改委、市公安局、市卫健委等76个部门、11个区、176个街道、2 790个村委、25家医院，累计节约了十余万小时的基层人力消耗，助力实现一个月左右的将本土每日新增病例控制在个位数，3个月左右本土每日新增病例基本为零。截至2020年7月21日，通过该系统发现并阻断感染人52人，累计“红码”人员305 024人，集中隔离11 310人，居家隔离1 095人，直接减少经济损失3 400万元；在后续企业复工复产工作中，系统进一步整合了全市企业数据，支持了对全市30 933家重点企业和579个重点项目的精准帮扶，有力推动了经济的快速复苏。

在平台的建设和应用过程中，X市数字政府有关部门也发现了一些不足的地方：一是个别政府部门信息化建设相对薄弱，

部分重要数据仍然通过电子表格的形式进行采集，影响了数据比对、清洗和分析的效率；二是对人工智能技术的利用尚不够深入。下一步X市将全面加强疫情防控相关信息化建设工作，全面汇聚政务信息资源，充分利用最新人工智能技术，为疫情防控常态化提供更加有力的支撑。

参考文献:

[1] 王琢, 荀亚玲, 张继福. 一种基于*K*-means的关联规则聚类算法[J]. 太原科技大学学报, 2016, 37(6): 429-437.
WANG Z, XUN Y L, ZHANG J F. An association rule clustering algorithm based on *K*-means[J]. Journal of Taiyuan University of Science and Technology, 2016, 37(6): 429-437.

[2] 李雷, 崔岩. 基于模糊聚类的改进的模糊关联规则挖掘算法[J]. 计算机技术与发展, 2012, 22(11): 18-21, 26.
LI L, CUI Y. An improvement of fuzzy association rules mining algorithm based on fuzzy clustering[J]. Computer Technology and Development, 2012, 22(11): 18-21, 26.

[3] 杨立波. 基于聚类的关联规则挖掘算法[J]. 太原大学学报, 2011, 12(3): 113-116.
YANG L B. Association rule algorithm based on cluster[J]. Journal of Taiyuan University, 2011, 12(3): 113-116.

[4] 胡庆辉, 丁立新, 陆玉靖, 等. 一种快速、鲁棒的有限高斯混合模型聚类算法[J]. 计算机科学, 2013, 40(8): 191-195.
HU Q H, DING L X, LU Y J, et al. Rapid robust clustering algorithm for Gaussian finite mixture model[J]. Computer Science, 2013, 40(8): 191-195.

[5] 王文明, 谭毓安. 基于EM聚类算法的机器人视觉场景深度分类方法[J]. 信息网络安全, 2013(6): 54-60.
WANG W M, TAN Y A. TA classification method of robot visual scene depth based on EM clustering algorithm[J]. Netinfo Security, 2013(6): 54-60.

[6] 曹家庆, 吴观茂. 基于MapReduce的分布式贪心EM算法[J]. 信息技术与网络安全, 2018, 37(5): 84-87, 92.
CAO J Q, WU G M. Greedy EM algorithm based on MapReduce framework[J]. Information Technology and Network Security, 2018, 37(5): 84-87, 92.

[7] 夏棒, EMILION R, 王惠文. Dirichlet混合样本的EM算法与动态聚类算法比较[J]. 北京航空航天大学学报, 2019, 45(9): 1805-1811.
XIA B, EMILION R, WANG H W. A comparison between EM algorithm and dynamical clustering for Dirichlet mixture samples[J]. Journal of Beijing University of Aeronautics and Astronautics, 2019, 45(9): 1805-1811.

[8] 文晓艺, 郝程程. 基于奇异值分解的新闻标题聚类研究[J]. 计算机技术与发展, 2020, 30(2): 42-46.
WEN X Y, HAO C C. Study on news header clustering based on singular value decomposition[J]. Computer Technology and Development, 2020, 30(2): 42-46.

[9] XIAO D L, ZHANG H, LI Y K, et al. ERNIE-GEN: an enhanced multi-flow pre-training and fine-tuning framework for natural language generation[J]. arXiv preprint, 2020, arXiv:2001.11314.

[10] 杜秀英. 基于聚类与语义相似分析的多文本自动摘要方法[J]. 情报杂志, 2017, 36(6): 167-172.
DU X Y. Multi-document automatic summarization based on clustering and semantic similarity analysis on cloud computing platform[J]. Journal of Intelligence, 2017, 36(6): 167-172.

[11] 齐小英. 基于NLPIR的人工智能新闻事件的语义智能分析[J]. 信息与电脑(理论版), 2019, 31(20): 104-107.
QI X Y. Semantic intelligence analysis of artificial intelligence news events based on NLPIR[J]. China Computer & Communication, 2019, 31(20): 104-107.

作者简介

李刚（1978– ），男，博士，广州市数字政府运营中心正高级工程师、主任，主要研究方向为数字政府、网络安全等。

郑佳（1981– ），女，广州市数字政府运营中心副主任，主要研究方向为数字政府。

尹华山（1977– ），男，广州市数字政府运营中心高级工程师，主要研究方向为数字政府。

黄文超（1979– ），男，广州市数字政府运营中心工程师，主要研究方向为数字政府。

收稿日期： 2020-08-03
通信作者： 尹华山，yinhs@gz.gov.cn