文章编号:1671-9352(2024)03-0095-12 DOI:10.6040/j.issn.1671-9352.7.2023.2681

政府开放数据中个人信息披露识别与统计方法

陈海粟,廖佳纯*,姚思诚
(南湖实验室大数据技术研究中心, 浙江 嘉兴 314002)

摘要:为推进数据开放过程中个人信息保护,深入分析政府开放数据中个人信息的披露现状:首先从相关平台中获取数据,并对其预处理,根据字段、表名等特征筛选出含有个人信息的数据;其次利用敏感信息识别方法识别数据中各类个人信息,并将其映射到个体,以统计个体数量同时检测其关联数据;最后通过数据可视化,直观展示个人信息披露现状。虽然部分公共数据开放平台虽然对公共数据进行分级分类以及去标识化等处理,但是已开放的数据中依旧包含大量直接展示的个人信息,需要在数据规范化分级分类、敏感信息识别和敏感信息脱敏等方面进行完善。

关键词:大数据隐私;个人信息;政府开放数据;信息识别;统计分析
中图分类号:TP391.1 **文献标志码**:A
引用格式:陈海粟,廖佳纯,姚思诚.政府开放数据中个人信息披露识别与统计方法[J]. 山东大学学报(理学版),2024,59(3):95-106.

Identification and statistical analysis methods of personal information disclosure in open government data

CHEN Haisu, LIAO Jiachun*, YAO Sicheng
(Research Center of Big Data Technology, Nanhu Laboratory, Jiaxing 314002, Zhejiang, China)

Abstract: To promote the protection of personal information during data opening, an in-depth analysis of the current status of disclosure of personal information in the open government data is conducted. Firstly, the paper obtains the datasets from relevant platforms and pre-process to classify the datasets that containing personal information based on features such as field and table names, etc. Then, methods of sensitive information identification are applied to identify and extract various types of personal information in the data, and map the information back to individuals to summarise the total number of individuals and detect their associated data. Through data visualizations, the current status of personal information disclosure could be examined. Although some open government data platforms may have implemented certain measures such as data categorization and de-identification, the published open datasets still contain a large amount of personal information, which is required to be improved in terms of data categorization and classification, sensitive information identification and data desensitization in a normative and accurate manner.

Key words: big data privacy; personal information; open government data; information identification; statistical analysis

0 引言

在全球信息技术迅猛发展的大背景下,数据已成为国家重要的基础性战略资源。大数据正引领新一轮科技创新和产业格局变革。我国在国家层面明确提出要培育数据要素市场,而该过程的探索离不开数据资源的整合、共享、开放、开发和利用[1]。2015 年 8 月,国务院发布《促进大数据发展行动纲要》,明确要求加快政府数据开放共享、推动资源整合、提升治理能力,提出中国要在 2018 年底前建成国家政府数据统一开放平

收稿日期:2023-04-29; **网络出版时间**:2023-12-14 11:16:31
网络出版地址:https://link.cnki.net/urlid/37.1389.N.20231213.0955.002
基金项目:南湖实验室小微课题资助项目(NSS2023C2002)
第一作者:陈海粟(1999—),男,硕士,研究方向为信息处理、智慧城市与个人信息保护. E-mail:hschen@ nanhulab.ac.cn
*通信作者:廖佳纯(1989—),女,助理研究员,博士,研究方向为信息处理、智慧城市与个人信息保护. E-mail:jliao@ nanhulab.ac.cn

台[2]。政府部门在履行行政职能、管理社会公共事务的过程中采集和储存了大量数据。在保障国家秘密、商业秘密和个人隐私的前提下,如果将政府数据最大限度地开放出来,让社会进行充分融合和利用,合力构筑数据基础设施,有利于释放数据能量,激发创新活力,创造公共价值。公共数据是我国政府数据的典型代表,是国家相关机构在依法履职或提供公共服务过程中收集、产生的数据,截止2022年10月,我国各地地方政府已经建立并上线省级公共数据开放平台共21个(包括省和自治区,不包括直辖市和港澳台)、城市公共数据开放平台共187个(包括直辖市、副省级和地级行政区)[3]。这些平台不仅在推进政府部门数据共享、促进政府数据资源利用中发挥着巨大的作用,也服务于民众从出生到教育、就业、医疗等生产生活的方方面面。在数字化的进程中,大量个人信息中的隐私部分作为公共数据被数据化收集、掌握和处理,因此公共数据的开放必然涉及了个人信息保护的问题。

当前对政府数据开放平台的个人信息保护研究主要聚焦于法律、政策等宏观层面。例如采用层次分析法对平台隐私政策进行评估,具有一定的简明性和广泛的适用性[4]。然而这类方法依靠大量人工设定的维度与指标,有主观性色彩,在个人信息披露的风险刻画上不够深刻[5]。另一方面,仅对政策分析,而忽略实际数据内容本身也会使得评估结果不够直观,缺乏可信度。

对此,需要从开放的公共数据中有效提取个人信息,再利用这些信息进行针对性的个人信息披露分析,提高评估的客观性与可信度。广义上来说,个人信息也属于敏感信息的一类,对个人信息的识别工作也可以看成是敏感信息识别的一种。对于大型数据表的敏感性分析可以利用基于元数据的敏感数据识别方法进行识别。我国公共数据开放平台的数据来自数字化改革后的不同的信息系统,格式复杂,各类别的数据可能来自不同的政府部门,没有统一且严谨的数据结构约束,同时也缺乏规范的业务数据的字段名命名标准。仅仅利用元数据信息(如数据表名、字段名等)不能准确地完成信息识别的任务。

针对上述问题,本文做出的创新之处在于:针对不同的类别的个人信息,提出严谨的识别与评估方法,改善个人信息识别中的误判问题;提出一种基于数据挖掘的个人信息披露分析方法,面向政府数据平台开放的公共数据,从数据的角度全面、客观地展示平台中个人信息披露情况;面对复杂的数据环境,结合个人信息识别、分析工具,综合利用信息抽取、关联分析等方法,提供一套端到端的个人信息识别工具,为后续的数据脱敏等工作提供上游的技术保障,有效减少数据开放平台个人信息泄露情况,促进数据的安全流转与共享。

1 相关工作

1.1 基于关键词匹配的敏感信息识别方法

传统的敏感信息识别大多基于关键词匹配的方法,通过人工设定筛选出有关敏感信息的关键词进而组建词表,并事先设定好判断待测文本中是否存在敏感信息的阈值,再利用多模匹配算法结合词表与阈值进行敏感性检测。

在敏感信息中一部分为编码类信息,其数据内容为一串统一编码的数字或字母组合的信息。对于个人信息而言,这类信息可以是身份证号、银行卡号、邮箱号等。由于这些信息都有明确的编码规则,因此可以编写识别的规则模板或建立特征值集合,进行快速检索。例如按身份证号码的特征组合码,建立特征值集合,再设置匹配算法,从而在数据库中识别身份证号信息[6]。另外,对于有格式要求的信息,例如邮箱号等,也可以利用其格式要求编写正则表达式对数据进行验证,并提取出通过验证的信息。

1.2 基于命名实体识别的敏感信息识别

敏感信息识别中另一种常用的技术是命名实体识别。该方法可以将敏感信息定义为不同的实体,并将其标注出来。个人信息里的个人姓名、住址、单位信息等都属于命名实体识别中的命名实体大类。对这些个人信息的识别也可以看成是命名实体识别任务。主要的方法包括基于规则、基于统计、基于深度学习等。早期主要利用基于规则的方法,例如在一个姓名语料库中统计出姓氏的使用频率、姓名首字的使用频率、姓名尾字的使用频率,以此构建人名用字知识库与人名规则,通过这个规则能够在一段文本中识别出人名信息[7]。

基于规则方法能有效地识别到相应的信息,但也有着泛化能力差、规则编写成本高的问题。因此,也有较多的学者采用基于统计与机器学习的方法,例如利用隐马尔可夫模型对一段文本进行标注,得到文本中的

实体类别信息,再从中提取需要的信息[8]。此类模型不需要繁琐的规则设计,同时也能针对特定的领域进行训练,有着更好的泛化能力。

随着计算机技术的发展,深度学习由于其强大的特征表示能力受到了空前关注,基于深度学习的命名实体识别模型在识别效果上也有较大提升。Guillaume 提出结合 Bi-LSTM 与条件随机场的模型。该模型利用 Bi-LSTM 对文本的信息进行编码,再利用条件随机场进行解码,得到最终的序列标注信息,相比基于机器学习的模型在识别的效果有较大提升[9]。

1.3　政府数据开放平台个人信息保护研究

目前国内外学者主要基于隐私政策角度,对不同的政府数据平台个人信息保护情况进行评估。例如采用内容分析法对《中华人民共和国个人信息保护法》进行分析概括,确立影响因素与分析维度,之后采用层次分析法构建指标体系,再运用德尔菲法对不同的指标权重打分,最后采用帕累托分类法区别地分析各级指标,计算得到政府数据平台的评价分值[10]。与此类似,杜荷花基于隐私政策视角,对美英奥政府数据开放平台隐私政策进行分析,从政府义务告知、隐私安全保护管理、个人权利保障 3 个维度构建评价指标体系,并采用该指标体系评估我国政府数据开放平台的隐私保护[11]。

上述工作大多都通过分析政策与法规量化数据开放平台的个人信息保护情况,缺少基于数据内容的直观分析与展示,特别是对个人信息披露的风险刻画不够客观。对此,一些研究通过 *K*-匿名评估一个数据集去识别化的程度,从而刻画一个数据集的个人信息披露风险[12];也有学者基于隐私保护数据挖掘,提出一种将多个去标识化的政府开放数据集成,再进行数据挖掘的方法,以此衡量隐私披露风险程度[13]。

由此可见,政府数据开放平台的个人信息保护工作已经得到一定的关注,但研究成果仍未形成体系,尤其是针对数据内容的统计分析较少。因此,本文将通过深入分析某地级市的公共数据开放平台中个人信息的披露现状,剖析我国在开放政府数据的个人信息保护方面尚存在的问题,响应“形成政府、企业、相关社会组织、公众共同参与个人信息保护的良好环境”的号召,为实现二十大报告中“提高公共安全治理水平,加强个人信息保护”贡献力量。

2　方法论

本研究将在一套识别和统计的算法应用框架中,以真实的公共数据开放平台作为实例,具体地对数据集中隐私信息进行识别和统计并对平台的个人信息披露情况展开评估。

应用框架中信息识别为人机交互式,信息识别的机器识别部分依赖正则表达式和命名体识别技术,人工部分对识别结果进行辅助性判断。识别完成后识别结果的统计分析在两个尺度下进行,为平台内每个数据集内直接披露情况的统计和关联平台内多个数据集导致的披露情况的统计。

本研究提出的应用框架如图 1 所示,主要分为数据源数据集获取、个人信息数据集筛选、个人信息识别和分析统计 4 个子模块。

2.1　基础术语

在具体阐述子模块方法之前,为描述公共数据的结构特点与内容的随机性,本文先对公共数据进行建模并依据 GB/T 37964—2019《信息安全技术—个人信息去标识化指南》,对其中给出个人信息相关的基本术语和定义进行建模[14]。

定义 1　常规的公共数据表由一组指示数据表的全局信息和一个与之对应的结构化数据部分构成。称 $(\boldsymbol{e},Q,R)$ 为一个数据表的基础数据结构,其中 $\boldsymbol{e}$ 为关于该数据表的如数据表名称、所属领域、来源部门等所有全局信息构成的向量,Q 为该数据表的结构化数据部分中的字段向量集合,R 为 D 的结构化数据部分中的记录行向量集合。

设公共数据表集合为 δ, 给定任意公共数据表 $D\in\delta$,设 $(\boldsymbol{e},Q,R)$ 为 D 的基础数据结构,

$$\boldsymbol{e}=[\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_l], \tag{1}$$

式中 l 为关于 D 的全局信息的信息数量, 任意 $\varepsilon\in\boldsymbol{e}$ 为关于 D 的任意全局信息;

$$Q=\{\boldsymbol{q}_1,\boldsymbol{q}_2,\cdots,\boldsymbol{q}_m\}, \tag{2}$$

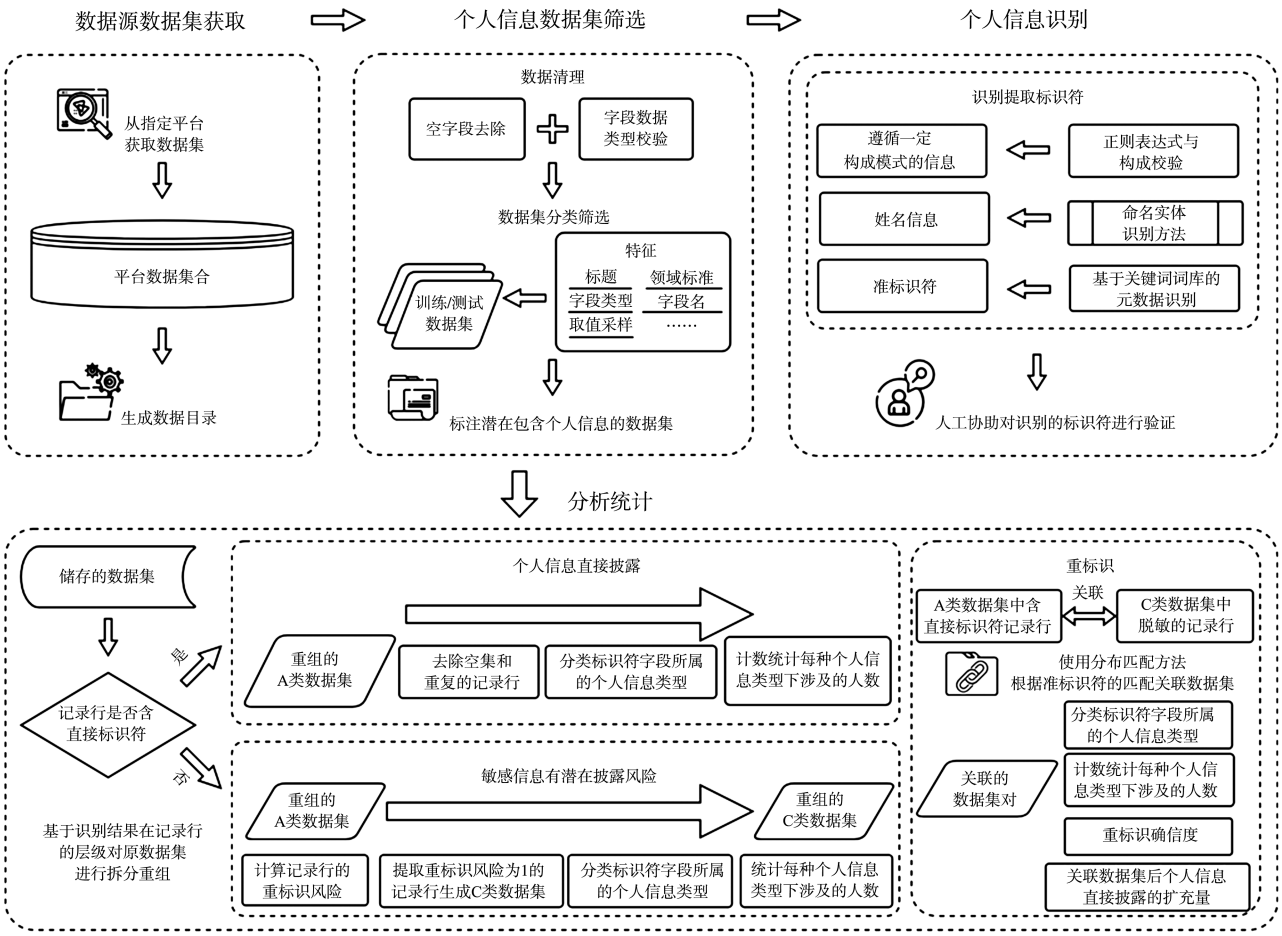

图 1 应用框架基本架构

Fig.1 Application framework

对于 $i=1,2,\cdots,m$, $\boldsymbol{q}_i \in Q$,有

$$\boldsymbol{q}_i=[d_{(i,1)},d_{(i,2)},\cdots,d_{(i,n)}], \tag{3}$$

式中 $\boldsymbol{q}$ 为字段向量集合 Q 中任意字段,d 为结构化数据部分中储存在单元格中的值,m 为 Q 的长度(元素数),n 为记录行向量集合 R 的长度(元素数),则 Q 可描述为其包含 m 个字段向量,每个字段向量由 n 个单元格的值表示;

$$R=\{\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_n\}, \tag{4}$$

对于 $j=1,2,\cdots,n$, $\boldsymbol{r}_j \in R$,有

$$\boldsymbol{r}_j=[d_{(1,j)},d_{(2,j)},\cdots,d_{(m,j)}], \tag{5}$$

式中 $\boldsymbol{r}$ 为 R 中任意记录行向量,则 R 可描述为其包含 n 个记录行向量,每个记录行向量由 m 个单元格的值表示。特别地,当 $j=1$ 时,$\boldsymbol{r}_1$ 可表示为 Q 中每个字段向量的字段名信息所构成的向量。

定义 2 对于任意字段向量 $\boldsymbol{q} \in Q$,其中非空元素可能存在取值重复。可以对任意字段向量 $\boldsymbol{q}$ 的取值元素进行去重操作,获得其唯一取值集合。称去重操作为 τ, 定义函数 $v:\boldsymbol{q}\to p$,表示去重操作 τ 作用至任意字段向量 $\boldsymbol{q}$ 产生其唯一取值集合 p 的过程。对于 $i=1,2,\cdots,m$, $\boldsymbol{q}_i \in Q$,有 $v(\boldsymbol{q}_i,\tau)=p_i$。定义函数 $R:Q\to P$,表示为对字段向量集合 Q 中所有向量元素均取其唯一取值集合构成新集合 P 的过程,有

$$P=R(Q)=\{p_1,p_2,\cdots,p_m\}。 \tag{6}$$

定义 3 设 θ 为字段唯一值取值占比,指示字段向量 $\boldsymbol{q}$ 中所有非空元素的唯一性的分布特征。对于 $i=1,2,\cdots,m$, $\boldsymbol{q}_i \in Q$,均有

$$\boldsymbol{q}_i^{*}=[d_{(i,j)} \mid d_{(i,j)} \neq \varnothing,\ j=1,2,\cdots,n], \tag{7}$$

$$\theta_i=\frac{L(p_i^{*})}{L(\boldsymbol{q}_i^{*})}, \tag{8}$$

式中 $\boldsymbol{q}_i^*$ 为字段向量 $\boldsymbol{q}_i$ 中所有非空元素重新构成的新向量,p_i^* 为向量 $\boldsymbol{q}_i^*$ 的唯一取值集合,$L(\boldsymbol{q}_i^*)$ 为向量 $\boldsymbol{q}_i^*$ 的长度(元素数),$L(p_i^*)$ 为唯一取值集合 p_i^* 的长度(元素数)。

定义 4　微数据是一个结构化的数据表,其中每条记录行对应一个个人信息主体,每个字段对应一个属性。称$(\boldsymbol{e},M,N)$为一个数据表中的微数据结构,其中 $\boldsymbol{e}$ 为该数据表的所有全局信息构成的向量,M 为该数据表的微数据结构中的字段向量集合,N 为该数据表的微数据结构中的记录行向量集合。

给定任意公共数据表 $D\in\delta$,假设 D 中存在微数据结构,则设$(\boldsymbol{e},M,N)$为 D 中存在的微数据结构,其中,$M\subseteq Q$, $N\subseteq R$,需要同时满足以下条件:

1) $M\neq\varnothing$,任意字段向量 $\boldsymbol{s}\in M$ 可对应一个属性;

2) $N\neq\varnothing$,任意记录行向量 $\boldsymbol{t}\in N$ 可对应一个个人信息主体;

定义 5　标识符是微数据中一个或多个属性,可以实现对个人信息主体的唯一识别,分为直接标识符和准标识符。对于 D 中的微数据结构$(\boldsymbol{e},M,N)$,存在标识符集合 $F=\{f_1,f_2\}$,其中$f_1\in F$ 为直接标识符集合,$f_2\in F$为准标识符集合。

对于任意字段向量 $\boldsymbol{s}\in M$,记 $L(\boldsymbol{s})$ 为字段向量 $\boldsymbol{s}$ 的长度(元素数),若满足 $\sum\boldsymbol{s}_x\odot d=1$,其中,$x=1,2,\cdots,L(\boldsymbol{s})$, $d\in s$,$\odot$表示同或运算符号, 则 $\boldsymbol{s}\in f_1$,可解释为字段向量 $\boldsymbol{s}$ 为直接标识符属性,可以在特定环境下单独对微结构数据$(\boldsymbol{e},M,N)$中记录行向量集合 N 所对应每个个人信息主体均实现唯一识别,常见的直接标识符有姓名、身份证、手机号码等;若不满足,则 $\boldsymbol{s}\in f_2$,可解释为字段向量 $\boldsymbol{s}$ 为准标识符属性,需要结合其他准标识符属性才有可能实现对微结构数据$(\boldsymbol{e},M,N)$中记录行向量集合 N 所对应每个个人信息主体实现唯一识别,常见的准标识符有性别、职业、学历等。

定义 6　等价类是指对于特定行,在微数据中与其准标识符属性相同值的所有数据记录行。假设对于 D 中的微数据结构$(\boldsymbol{e},M,N)$中的标识符集合 F,其中存在准标识符集合$f_2=\{\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_\gamma\}$, $\gamma\leqslant m$, $\boldsymbol{s}_\gamma\in M$,记 $L(N)$为记录行向量集合 N 的长度(元素数),则对于 $y=1,2,\cdots,L(N)$,均有

$$\zeta_y=\{\boldsymbol{s}_x[y],\ x=1,2,\cdots,\gamma\},\tag{9}$$

其中,ζ_y 表示为 D 中的微数据结构$(\boldsymbol{e},M,N)$的任意记录行 y 在所有准标识符属性下的取值所构成的集合。

对任意记录行 y 获取其等价类,有

$$\sigma_y=\{k\mid\zeta_k=\zeta_y,\ k=1,2,\cdots,L(N)\},\tag{10}$$

其中,σ_y 表示为记录行 y 与在所有准标识符属性下的取值均相同的其他所有记录行的索引集合,即集合内任意记录行 y 的等价类。$L(\zeta_y)$表示为该记录行 y 构成的向量的长度(元素数),即等价类维度。$L(\sigma_y)$表示为该记录行 y 所形成的等价类的长度(元素数),即等价类大小。

定义 7　重标识指将去标识化的数据表重新关联到原始个人信息主体或一组个人信息主体的过程。给定 D 中的微数据结构$(\boldsymbol{e},M,N)$,若其进行了去标识化处理,则该微数据结构中每一行都存在重标识的概率。

对于特定行,重标识概率取决于该微数据结构中该行形成的等价类的大小。由于同一等价类中各记录行不可区分,对于同一等价类下各记录行均存在相同的被重标识的概率, 故对于记录行 y 所对应的个人信息主体仅存在被重标识的概率,为 $\mu_y=\dfrac{1}{L(\sigma_y)}$,其中,$\mu_y$ 表示为该记录行 y 的重标识风险。对于该微数据结构$(\boldsymbol{e},M,N)$,存在重标识概率最大值,为 $\beta=\max\limits_{y=1,2,\cdots,L(N)}(\mu_y)$。

2.2　数据源数据集获取

本研究将对某平台上不同领域的所有开放数据进行下载。公共数据开放平台为广大人民群众提供政务服务,其面向社会大众,具有开放性和公开性。在平台实际使用过程中,任何拥有个人身份证或统一社会信用代码的个人或法人可以注册账号,登录下载平台上的任意无条件开放数据,供个人使用或分享给其他个人或组织。对在平台上实际下载获取到的所有有效的无条件开放数据集进行数据集信息汇总并制作数据集目录。所获取的无条件开放数据均为 Excel 格式的结构化数据集,其字段取值中随机嵌套非结构化的描述性文本。以该形式完成的数据集获取可以看作对平台无条件开放的数据集进行数据快照,即记录在数据集获取时某平台的开放数据状态。后续的分析和统计均基于该数据快照,不考虑平台后续对开放数据的更新。

2.3 个人信息数据集筛选

该模块包含两子部分:数据清理和数据集分类。鉴于所获取的开放数据来自不同部分,其数据质量往往未经检查。预处理的第一步为数据清理,即对数据集中完全为空值的无效字段进行清除并对错误数据类型的字段进行数据类型校正。无效字段和错误数据类型的字段对于分类模型训练和后续的信息识别和统计分析等任务而言均具有干扰作用。在数据清理完成后,需要进行数据集分类以分类筛选潜在包含个人隐私信息的数据集。个人信息数据集分类筛选是识别检测数据集个人信息披露情况的基础任务,仅有在确保数据集的内容与个人信息相关后才能开展后续相关处理任务。分类筛选以数据集中每个字段为单位,使用的相关特征信息包含平台发布数据集时对数据集所属领域的描述、数据集的标题、字段名、字段唯一值取值占比、字段样本取值的各项特征(如样本的取值是否为中文、数字和英文等)。在对可用相关特征信息完成向量化转化后,输入至预训练的机器学习分类模型,对数据集的各字段完成初步的分类标注,并初步确认数据集的标注,即是否潜在包含个人信息。随后进行人工排查,对标注为个人信息数据集的数据集进行人工核验。

2.4 个人信息识别

在该实例分析的展示中,本研究着重识别敏感信息中的个人基本信息、个人身份信息、个人生理健康信息、个人工作教育信息、个人财产信息和其他个人信息共六种个人信息。划分标准参照《网络安全标准实践指南—网络数据分级分类指引》中附录 B 表 B.1[15]。

个人信息识别包含直接标识符的识别和准标识符的识别。在本研究中,直接标识符的识别方法主要分为两种。第一种为正则表达式,是由字符序列定义的搜索模式,主要应用在搜索严格遵循一定构成模式的数据,如身份证、银行卡、手机号、邮箱地址等;第二种为基于深度学习的命名实体识别方法,主要应用在对文本序列中的姓名的识别和提取[16]。根据字段的数据组成类型,即字段的取值构成为纯数字、数字与字符串文本混合、纯字符串文本、长短文本等,对于不同的字段数据组成类型采取不同的识别方法组合策略对直接标识符进行识别。此外,为提高识别效率,识别将在字段唯一取值集合中进行,并通过建立从字段序列唯一取值元素回到原字段序列的映射字典的方式,将识别结果映射回原字段序列。准标识符的识别方法主要为基于关键词词库的元数据识别。由于各准标识符类型复杂,往往没有一个标准的构成模式,而在结构化的数据集中,结构本身已经传递一些信息,对数据集的字段名集合中通过使用关键词词库进行信息类型的检索可以判别相应的准标识符。在某平台中所获取的数据集里,匿名化的数据往往包含特殊字符“ * ”,通过该构成模式可以识别出已进行去标识化的个人信息。在敏感信息的识别过程中同步进行敏感信息的提取,以便进行识别结果的审查和后续的统计。机器部分的识别完成后,由人工对机器识别结果通过采样进行辅助检查。

2.5 分析统计

依据直接标识符的识别结果,即数据集中记录行含直接标识符信息的有无,可以对原数据集在记录行的层级进行拆分重组,重组生成每行均直接披露敏感信息的 A 类数据集和每行不直接披露敏感信息而仅有披露风险的 B 类数据集。

2.5.1 单个数据集内直接披露情况统计

对 A 类数据集中数据表各字段取唯一取值集合,然后对数据集涉及的直接标识符字段和准标识符字段进行上述提及的六种个人信息类型的判定,直接标识符字段的个人信息类型的分类判定依据其识别的标注结果,准标识符字段的个人信息类型的分类判定主要采用基于关键词词库的元数据识别。在各标识符字段完成个人信息类型判定的基础上,分别对 A 类和 B 类数据集在各个人信息类型分类下敏感信息直接披露涉及的人数进行统计:

1) 在 A 类数据集中,对各标识符字段下记录行的数量进行计数统计,将记录行计数统计结果作为该标识符字段下敏感信息直接披露涉及的人数的指示,再依据各标识符字段所属的个人信息类型分类,统计各个人信息类型分类下敏感信息直接披露涉及的人数。

2) 在 B 类数据集中,对各记录行计算重标识风险。数据集内记录行的重标识风险为 1 意味着该记录行对应的个体在该数据集环境下可以被唯一定位。抽取 B 类数据集中重标识风险为 1 的记录行,重组生成 C 类数据集。

3）在 C 类数据集中采取和 A 类数据集类似的方法,对各个人信息类型分类下敏感信息直接披露涉及的人数进行计数统计。

2.5.2　关联多个数据集的信息披露情况统计

本研究采用分步匹配方法,以准标识符的匹配为参考基准,将 A 类数据集中含直接标识符信息的记录行和 C 类数据集中去标识化的记录行进行数据集关联:

1）取 A 类数据集中的任意数据集 A_i 和 C 类数据集中的任意数据集 C_j,分别获取两个数据集的准标识符字段集合,逐一检视两数据集中各对准标识符字段的字段名及取值,将包含同一种个人信息类型且有共同取值的字段相配对,获得两数据集中所有的可匹配准标识符字段对。

2）对两数据集记录行的对应字段值做匹配分析。依据步骤一所确定的两数据集准标识符字段对的取值,对两数据集的记录行逐一分析,仅将所有准标识符字段对取值均相同的记录行相匹配,若两记录行的所有准标识符字段对取值均匹配,则可以作出该对记录行对应同一个个人的判断。

在完成记录行关联匹配后,成功匹配的准标识符数量可以用于衡量匹配记录行对应同一个体的可信度。对于配对数据集 A_i 和 C_j, 数据集 A_i 扩充的信息量为数据集 C_j 的准标识符数量减去两数据集中所有可匹配的准标识符字段对的数量。此外,鉴于存在 A 类数据集中 A_i 的一条记录行通过分步匹配方法后可能与 C 类数据集中 C_j 的 α 条重标识风险为 1 的记录行在所有可匹配准标识符字段对的取值上均匹配,则定义,对于配对的数据集 A_i 和 C_j,数据集 A_i 在关联匹配后扩充的个人信息的确信度为$\frac{1}{\alpha}$。最后依据关联匹配结果,统计在不同可信度和不同确信度下实现重标识的记录数目。

3　结果

本文对真实公共数据开放平台展开识别和统计的算法框架的应用,通过对平台敏感信息披露的识别与结果统计从而对平台的个人信息保护情况进行评估。

3.1　数据集

本文以某平台作为实例,截止 2022 年 11 月,下载并使用了平台当时发布的全部 339 个无条件共享数据集,经过分类筛选,有 94 个数据集包含个人信息。该 94 个数据集的数据容量(即单位数据集内的记录行数量)与属性数量(即单位数据集内的列数)如图 2 展示。

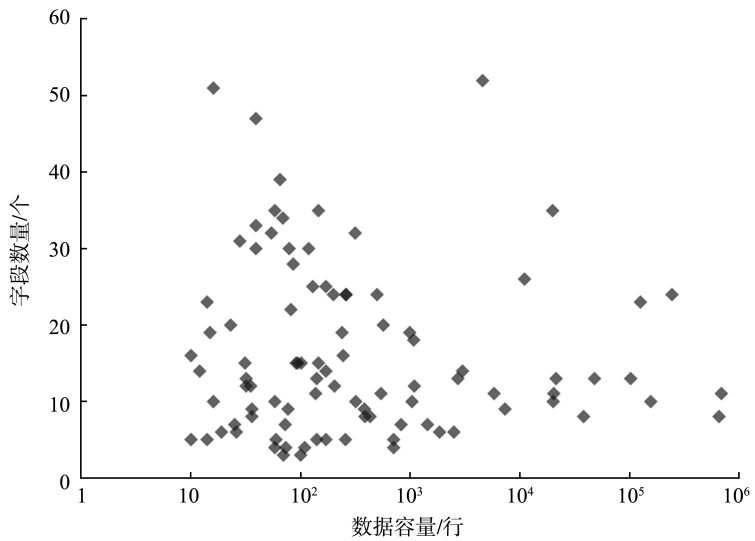

图 2　对象平台含个人信息数据集的数据容量和字段属性数量综合情况

Fig.2　Plot of data size and number of fields in the datasets from test platform

3.2　识别算法对比分析

为评估本文识别算法的性能,选取对象平台中较为典型的两个数据集进行实验,对其中涉及的人名、手机号、身份证号、车牌号、银行卡号信息进行识别,评估指标为漏检的信息数目与误检的信息数目。

3.2.1 评估数据集

在该平台上的数据集格式多样,有结构化的数据表,同时也有非结构化的数据表。样例数据展示原数据集部分记录情况,其记录内容在原数据集中均明文展示。为本文展示需要,明文展示的潜在涉及个人隐私的信息已进行脱敏处理,如表 1、2 所示。

表 1 婚姻矛盾纠纷数据集样例(总计 137 条)

Table 1 Samples for the dataset of Marital Conflict (137 records in total)

主要诉求	调处情况
刘 * 与丈夫陈 ** 2010 年结婚,育有 2 个孩子……	联系 ** 派出所询问情况,请派出所帮助开具家暴告诫书……

表 2 老人信息数据集样例(总计 4 619 条)

Table 2 Samples for the dataset of the Elders Information (4 619 records in total)

门磁 ID	老人证件号码	老人姓名
8632170 **** 4067	****** 1939 ****** 47	许 **

3.2.2 对比算法

本次实验采用的基线识别算法具体识别模式见表 3,描述如下:

1) 统一编码类信息:对严格遵循一定构成模式的数据,如手机号、身份证号、车牌号、银行卡号等,基于目标类信息常见的匹配模板,编写正则表达式进行识别。

2) 姓名信息:利用联合的词法分析工具 LAC 对数据表字段进行命名实体识别,提取其中的姓名[16]。

表 3 基线识别算法罗列

Table 3 Listing of baseline algorithms for identification

识别类型	目标类信息	识别模式
统一编码类信息	身份证	正则表达式: r'^[1-9]\d{5}(18|19|([23]\d))\d{2}((0[1-9])|(10|11|12))(([0-2][1-9])|10|20|30|31)\d{3}[0-9Xx]$'
	手机号	正则表达式: r'^((\+? [0-9]{1,4})|(\(\+86\)))? (13[0-9]|14[57]|15[012356789]|17[03678]|18[0-9])\d{8}$'
	车牌号	正则表达式: r'^[京津沪渝冀豫云辽黑湘皖鲁新苏浙赣鄂桂甘晋蒙陕吉闽贵粤青藏川宁琼使领][A-HJ-NP-Z](?:((\d{5}[A-HJK])|([A-HJK][A-HJ-NP-Z0-9][0-9]{4}))|[A-HJ-NP-Z0-9]{4}[A-HJ-NP-Z0-9 挂学警港澳])$'
	银行卡号	正则表达式: r'(? <! [0-9a-zA-Z\-])[1-9](?:\d{11,18})(?! [0-9a-zA-Z\-])'
姓名信息	姓名	LAC 工具命名实体识别

本文使用的识别算法发展自基线算法,具体识别模式见表 4,描述如下:

1) 统一编码类信息:对于严格遵循一定构成模式的数据,参考实际公共数据中非结构化文本中敏感信息的识别情况,在常规的匹配模板的基础上编写更加精准的正则表达式,并同时对识别的具体数据如身份证、银行卡号进行合规校验。

2) 姓名信息:使用 HanLP 对数据表字段进行命名实体识别,提取其中的姓名[17]。

表 4 本文识别算法罗列

Table 4 Listing of developed algorithms for identification

识别类型	目标类信息	识别模式
统一编码类信息	身份证	正则表达式: r'[1-9]\d{5}(?:18|19|(?:[23]\d))\d{2}(?:(?:0[1-9])|(?:10|11|12))(?:(?:[0-2][1-9])|10|20|30|31)\d{3}[0-9Xx]' 身份证校验:模 11 算法校验、地区码和时间合规校验
	手机号	正则表达式: r'(? <! [0-9a-zA-Z\-])(?:\+? 86)? 1(?:(?:34[0-8])|(?:8\d{2})|(?:(?:[35][0-35-9]|4[14-9]|6[567]|7[0-8]|9[12389])\d))\d{7}(?! [0-9a-zA-Z\-])'

续表

识别类型	目标类信息	识别模式
统一编码类信息	车牌号	正则表达式: r'(? <! [锅容管瓶梯起索游车]\d{2}(? =[京津沪渝冀豫云辽黑湘皖鲁新苏浙赣鄂桂甘晋蒙陕吉闽贵粤青藏川宁台琼使领军北南成广沈济空海]{1}[A-Z]{1}[A-Z0-9]{4}(?:[A-Z0-9 挂领学警港澳]{1}|[A-Z0-9]{2}\(\d{2}\))))[京津沪渝冀豫云辽黑湘皖鲁新苏浙赣鄂桂甘晋蒙陕吉闽贵粤青藏川宁台琼使领军北南成广沈济空海]{1}[A-Z]{1}[A-Z0-9]{4}(?:[A-Z0-9 挂领学警港澳]{1}|[A-Z0-9]{2})(?! \d)'
	银行卡号	正则表达式: r'(? <! [0-9a-zA-Z\-])[1-9](?:\d{11,18})(?! [0-9a-zA-Z\-])' 银行卡校验:Luhn 规则校验和银行卡号前缀匹配
姓名信息	姓名	HanLP 工具命名实体识别

3.2.3　实验结果

由表 5、6 中可以看到,本文的识别算法在两类数据集上者有更少的漏检与误检,有助于提高后续的统计分析结果的可靠性。

表 5　婚姻家庭矛盾纠纷数据集识别结果

Table 5　Identification results for the dataset of Marital Conflict

	统一编码类信息(漏检/误检)	姓名(漏检/误检)
基线识别算法	0/4	10/20
本文识别算法	0/0	1/4

表 6　老人信息数据集识别结果

Table 6　Identification results for the dataset of the Elders Information

	统一编码类信息(漏检/误检)	姓名(漏检/误检)
基线识别算法	0/11	184/0
本文识别算法	0/0	0/0

基线识别算法在统一编码类信息上没有漏检,但有一定的误检,这是因为基线识别算法采用较为宽泛的匹配规则,这些规则是必要条件,因此并不会产生漏检。但该方法没有考虑到公共数据的数据结构的特点,因此容易误检编码类信息,例如将门磁 ID“8632170 * * * * 4067”中的“170 * * * * 4067”识别成手机号。本文的识别算法考虑了各类信息在数据集中出现的形式,对手机号这类无校验算法的信息采用前后的匹配限定,排除待匹配号码前后出现数字的情况,这个策略既能杜绝此类误检现象,又能避免漏检夹杂在中文里的手机号。

另一方面,本文综合考虑姓名信息出现的形式,使用更高效的命名实体识别方法[17],同时利用字段名、结构化程度、抽样识别率作为判断依据,召回满足条件的字段中所有的数据,帮助识别算法做到在非结构化数据中少误检,结构化字段少漏检,大大提高了识别的可靠性。

3.3　平台敏感信息披露的识别与统计结果

根据数据集敏感信息识别结果中记录行的直接标识符的有无,对该 94 个数据集进行拆分重组,其中有 8 个数据集所有记录行均含完整的直接标识符信息,故最终重组生成 60 个 A 类数据集和 86 个 B 类数据集。

3.3.1　单个数据集内直接披露情况展示

A 类数据集中所有记录行均包含直接标识符信息,通过对 A 类数据集的统计分析可以获取平台在单个数据集内各类型的敏感信息的直接披露情况。以数据集的领域标注和披露的个人信息种类为类别标签,A 类数据集披露的个人信息类型涉及人数如表 7 所示。

表 7 对象平台单个数据集内直接披露的各个人信息类型涉及人数情况

Table 7 The summarized number of related people with the direct disclosure of personal information in each single dataset from test platform

领域标注	披露的个人信息类型涉及人数/人					
	个人基本信息	个人身份信息	个人健康生理信息	个人教育工作信息	个人财产信息	其他个人信息
社会救助	42 111	24	19 413	0	8	0
市场监督	10 110	3	0	10 094	0	0
科技创新	2 965	0	0	2 965	0	0
气象服务	1 027	0	0	0	0	0
生态环境	1 260	0	0	667	0	0
生活服务	256	7	1	29	17	94
城建住房	159	0	0	0	0	0
教育文化	64 162	2 143	0	61 138	0	0
地理空间	123	0	0	0	0	0
交通运输	42	59	0	59	0	0
信用服务	399	19	0	0	0	0
机构团体	1	0	0	0	0	0
工业农业	16	0	0	0	0	0
其他	12	0	0	12	0	0

可以发现这 60 个 A 类数据集直接披露了超过 10 万人的个人基本信息,近 2 万人的个人健康生理信息以及超 7 万人的个人教育工作信息。通过检查具体披露的数据集后发现主要涉及人群为教育文化领域中如教师、学生、图书馆用户、志愿者等,社会救助领域中如残疾人、低保户与救助申请人等,以及市场监督领域中的监管管理人员、文件签发人、个体经营者等。

3.3.2 关联多个数据集的信息披露情况展示

将 C 类数据集中的含匿名化字段信息的记录行与 A 类数据集中完整披露直接标识符的记录行进行关联后,可能实现对重标识风险为 1 的已去标识化的记录行所对应个体的重标识,以及扩充包含个人可识别性信息的记录行所对应个人的信息种类。经过分步匹配方法,发现 A、C 数据集中能实现关联匹配的共有 27 组数据集对。其中有 9 738 个已存在个人信息直接披露的个人信息主体可以与平台提供的已进行去标识化操作但重标识风险为 1 的记录存在一定程度上的关联匹配。准标识符的匹配数量能够作为指示关联数据集重标识的可信程度的依据,其中姓氏的匹配也归在准标识符匹配数量中。关联后披露的个人信息及其涉及的人数如图 3 所示。

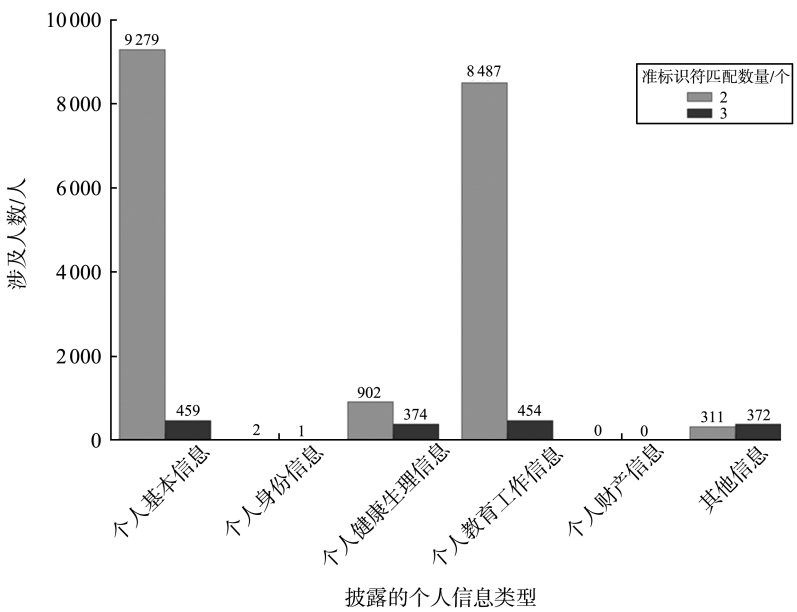

图 3 关联数据集进行重标识后披露的个人信息类型的涉及人数

Fig.3 The summarized number of related people with re-identification by data linking from test platform

通过关联数据集进行重标识,虽然平台披露的可识别个人的数目没有增加,但是关联成功后对于特定个人信息主体的信息数量会得到增长。依据 2.5.2 中提及的确信度概念,对关联匹配后扩增的信息量在不同确信度下进行了涉及人数的统计,如图 4 展示。

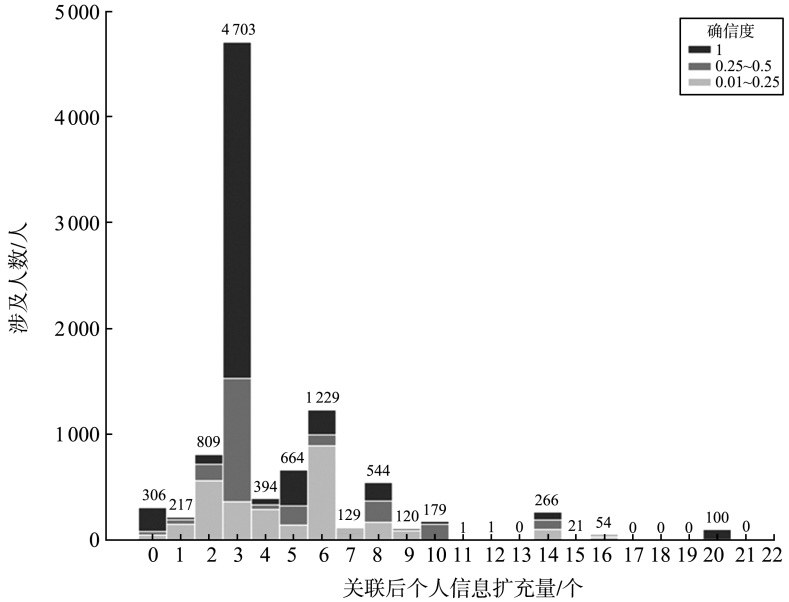

图 4　关联数据集进行重标识后在不同确信度下对个人信息主体个人信息的扩充量及其对应的涉及人数

Fig.4　The summarized number of related people with re-identification by data linking under certain degree of confidence and quantity of information extension from test platform

4　总结与展望

本文对我国政府数据开放平台的个人信息保护情况进行了探究,提出了一种评估开放数据中个人信息披露情况的方法框架。通过在某平台的实际应用,可以直观发现平台中存在大量可识别个人的敏感信息的直接披露与一定数量的关联数据集导致敏感信息被重标识的情况。当前进行的研究工作是对现存的迫切问题的一次探索性尝试,目前本文所展示的仅是探索工作的初步成果。未来的工作将围绕框架算法在分类和识别的精度、运行效率和识别覆盖面,以及以模块化的方式在算法框架中嵌入对识别到的个体进行去标识化的功能这多个方面进行进一步的开发和研究。

参考文献:

[1] 梅宏. 数据治理之路:贵州实践[M]. 北京:中国人民大学出版社, 2022:47.
MEI Hong. On data governance: practice in Guizhou[M]. Beijing: China Renmin University Press, 2022:47.

[2] 国务院. 国务院关于印发促进大数据发展行动纲要的通知[EB/OL]. (2015-09-05)[2023-02-12]. https://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm.
The State Council. Circular of the state council on printing and issuing the action outline for promoting the big data development [EB/OL]. (2015-09-05)[2023-02-12]. https://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm.

[3] 复旦大学数字与移动治理实验室. 中国地方政府数据开放报告—城市指数(2022 年度)[R/OL]. (2023-01-10)[2023-01-30]. http://ifopendata.fudan.edu.cn/report.
DMG Lab Fudan University. China's local government open data report—city index (2022)[R/OL]. (2023-01-10)[2023-01-30]. http://ifopendata.fudan.edu.cn/report

[4] 黄玥,周丽霞,蒲攀. 基于 AHP 方法的我国信息安全政策方案优化决策研究[J]. 现代情报, 2015, 35(3):77-81.
HUANG Yue, ZHOU Lixia, PU Pan. Study on the optimizing of information security policy based on AHP[J]. Journal of Modern Information, 2015, 35(3):77-81.

[5] 周林兴,周丽. 政府数据开放中的隐私信息治理研究[J]. 图书馆学研究, 2019(12):41-47.
ZHOU Linxing, ZHOU Li. Research on privacy information governance in open government data[J]. Research on Library

Science, 2019(12):41-47.

[6] 李立新,唐培洪,臧滔,等. 一种身份证号码识别方法、装置和电子设备:CN112380211A[P]. 2021-02-19.
LI Lixin, TANG Peihong, ZANG Tao, et al. The invention relates to a method, a device and an electronic device for the identification of resident identity card number: CN112380211A[P]. 2021-02-19.

[7] 闫萍. 基于规则和概率统计相结合的中文命名实体识别研究[J]. 计算机与数字工程, 2011, 39(9):88-91.
YAN Ping. Research on the identification for Chinese named entity based on combination of rules and statistic analysis[J]. Computer & Digital Engineering, 2011, 39(9):88-91.

[8] 俞鸿魁,张华平,刘群,等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报, 2006(2):87-94.
YU Hongkui, ZHANG Huaping, LIU Qun, et al. Chinese named entity identification using cascaded hidden Markov model [J]. Journal on Communications, 2006(2):87-94.

[9] GUILLAUME L, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: Association for Computational Linguistics, 2016:260-270.

[10] 孙瑞英,李杰茹. 我国政府数据开放平台个人隐私保护政策评价研究[J]. 图书情报工作, 2022, 66(12):3-16.
SUN Ruiying, LI Jieru. Research on the evaluation of personal privacy protection policies of government data open platforms in China[J]. Library and Information Service, 2022, 66(12):3-16.

[11] 杜荷花. 我国政府数据开放平台隐私保护评价体系构建研究[J]. 情报杂志,2020,39(3):172-179.
DU Hehua. On construction of privacy protection evaluation system of government data open platform in China[J]. Journal of Intelligence, 2020, 39(3):172-179.

[12] SWEENEY L. *K*-anonymity: a model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5):557-570.

[13] LEE J S, JUN S P. Privacy-preserving data mining for open government data from heterogeneous sources[J]. Government Information Quarterly, 2021, 38(1):101544.

[14] 全国信息安全标准化技术委员会. 信息安全技术—个人信息去标识化指南: GB/T 37964—2019[S]. 北京:中国标准出版社, 2019.
National Information Security Standardization Technical Committee. Information security technology—guide for de-identifying personal information: GB/T 37964—2019[S]. Beijing: Standards Press of China, 2019.

[15] 全国信息安全标准化技术委员会秘书处. 网络安全标准实践指南—网络数据分级分类指引[EB/OL]. (2021-12-31)[2023-01-30]. https://www.tc260.org.cn/upload/2021-12-31/1640948142376022576.pdf.
The Secretariat of National Information Security Standardization Technical Committee. Practice guide on network security standards—guidelines on classification of network data[EB/OL]. (2021-12-31)[2023-01-30]. https://www.tc260.org.cn/upload/2021-12-31/1640948142376022576.pdf.

[16] JIAO Zhenyu, SUN Shuqi, SUN Ke. Chinese lexical analysis with deep Bi-GRU-CRF network[EB/OL]. (2018-06-05)[2023-01-30]. https://doi.org/10.48550/arXiv.1807.01882.

[17] HE H, CHOI J D. The stem cell hypothesis: dilemma behind multi-task learning with transformer encoders[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominian: Association for Computational Linguistics, 2021:5555-5577.

(编辑:甄鹏)