政府数据中敏感数据识别与隐私计量研究*

臧国全[1,2]　王家振[1]　毕崇武[1,2]　耿瑞利[1,2]
[1] 郑州大学信息管理学院　郑州 450001　[2]郑州市数据科学研究中心　郑州 450001

摘　要：[目的/意义]通过分析政府数据隐私相关文本，设计敏感数据识别方案，构建隐私计量模型，计量敏感数据的隐私值，为政府数据隐私保护提供理论依据。[方法/过程]首先筛选政府数据隐私的相关文本构建样本库；然后依据文本的句法结构，抽取敏感数据项、核心动词、程度词、否定词等词汇，构建政府数据隐私语义词表；最后以上述词汇组成的敏感数据单元为基础，构建隐私计量模型。[结果/结论]该方法基于隐私相关文本，准确析出政府数据的敏感数据，客观计量政府数据对象的隐私值，可为政府数据的隐私风险防范及隐私保护规范化提供支持。

关键词：政府数据　数据隐私　个人隐私　语义词表　隐私计量
分类号：G250
DOI：10.13266/j.issn.0252-3116.2022.15.007

1　引言

政府数据是指政府部门在履行管理与服务职责过程中生产、采集和调用的数据，包括政府个人数据和政府公共数据。隐私是隐私主体（自然人）对其隐私客体（敏感数据内容）的敏感性和敏感程度的认知。因此，政府数据隐私仅存在于政府个人数据之中，是政府部门对公众个人数据进行采集、保存、处理、使用和开放过程中涉及的个人隐私。现代服务型政府的社会管理与服务需要庞大的数据支撑，这些数据涉及个人隐私越来越多，数据的利用与隐私保护之间的矛盾日益凸显。鉴于此，清晰识别数据隐私项目以及对隐私精确的计量，有助于制定有效的开放标准，规范隐私开放过程，防范隐私风险的产生。

个人隐私数据有多种类型且都具有敏感性，因此本文中统称为敏感数据。本文对政府部门采集的个人数据进行汇总整理，析出敏感数据词汇，结合法律法规文本、专家论著文本及隐私泄露案例文本，构建面向政府数据隐私的语义词表。抽取三个样本库中的敏感数据单元，构建基于模糊层次分析法的隐私计量模型，计量文本中隐私项的隐私值。通过客观视角下的隐私计量，为政府数据管理与开放过程中个人隐私保护的规范化与标准化提供理论基础与参考。

2　相关研究

本文对政府数据隐私保护实践的现状与隐私计量研究现状进行简要梳理，总结隐私保护困境与隐私计量研究的不足。选择基于政府数据隐私相关文本的客观计量，通过提出新的研究视角拟解决部分问题。针对政府数据隐私相关文本的处理主要为文本特征提取、文本的标签处理两方面内容。

2.1　政府数据隐私保护与计量相关研究

随着政府数据开放的兴起，个人隐私风险与保护问题便一直受到学者们的关注。针对我国的政府数据隐私保护实践中的问题也有着广泛讨论。在开放政策探讨方面，分别有学者从政府数据开放共享政策的相关文件[1]、各大城市的政府数据开放平台[2]、城市政府门户网站的“隐私声明”[3]等多角度出发调查政策制定到执行的全过程，认为地区间隐私保护政策差距较大，保护标准缺乏统一。在法律法规方面，我国目前仅有《政府信息公开条例》的制度建设性法规，政府主体对于个人信息的使用权限问题，即针对个人信息的使用规范，目前在制度层面尚无明确的应对[4]。此外，也有基于数据生命周期进行探讨的相关研究。分析政府

* 本文系国家社会科学基金重大项目“政府数据的隐私风险计量与保护机制创新研究”（项目编号：21&ZD338）研究成果之一。
作者简介：臧国全，院长，教授，博士生导师；王家振，硕士研究生；毕崇武，讲师，硕士生导师，通信作者，E-mail：767818984@qq.com；耿瑞利，副教授，硕士生导师。
收稿日期：2022-01-13　修回日期：2022-04-28　本文起止页码：66-75　本文责任编辑：易飞

数据采集、保存、使用和开放等各个阶段的隐私保护问题[5-6]，提出应从技术和管理相结合的视角完善法律体系，加强隐私保护相关标准规范约束等应对措施。

隐私计量是规范隐私保护制度、推动隐私分级保护的有效手段之一。常见的隐私计量是基于成本效益理论的价值计量，即视隐私为一种经济商品，依据隐私主体为保护个人数据隐私而产生的支付意愿，或牺牲个人数据隐私来换取货币补偿的接受意愿，计量隐私的货币价值。此类计量是基于隐私主体认知，所获得的结果主观性较强。在计量方法方面多采用实验经济学法：①联合分析法：如计量公众对于个人信息保护服务[7]或者患者对医疗数据保护服务的支付意愿[8]，通过探讨网上披露个人信息的成本和收益之间的权衡来计量接受意愿[9]；②离散选择法：如 S. Egelman 等[10]对用户位置数据和通话记录的支付意愿价格进行了计量，邓胜利等[11]通过分析不同类型的个人信息泄露与个体受偿意愿之间的关系对信息的价值进行了计量；③条件价值法：如黄逸珺等[12]调查得出了用户在牺牲个人信息的前提下愿意接受的补偿价格，V. Benndorf 等[13]评估了用户贡献给社交网络的个人信息隐私的货币价值，张凯亮等[14]测度了用户在一定隐私泄露概率情境下对其个人数据价值的认知以及个人数据隐私在泄露概率情境下的价值；④拍卖实验法：如通过测量用户为保护社交媒体上个人信息的付费意愿[15]或分享个人信息所需的最低金额[16]来计量个人信息的价值。

2.2 政府数据文本处理相关研究

文本特征提取常被用于政策类文本的解读与挖掘[17-18]，能够有效提取出政策核心。因此本文采取中文文本的特征提取方法，以准确提取政府数据隐私相关文本中的核心词汇，构建主题词表。特征提取的方法主要有三种：①语言学方法。即先对文本进行语言特征分析，再进行文本处理。主要是通过识别词间关系完成特征抽取。X. F. Li 等[19]提出通过分析基本名词短语(baseNP)与基本动词短语(baseVP)的特征，实现基于词组与关键词的中文文本特征提取；唐晓波等[20]基于依存句法分析构建了文本网络；葛斌等[21]提出一种多层最大熵模型，并结合语句特征与句间关系实现了句子主干词分析；涂安龙[22]使用 RS 语义分析模型对筛选后的特征空间进行了特征提取。②统计学方法。指对文本进行集中分析过程中，发现并识别文本语言特征。曹钲晨[23]使用 Kmeans＋获得聚类中心，统计了文章中每个句式所归属的簇；E. A. Dietz 等[24]基于 WordNet 中概念的相似度来形成概念聚类的类别；安亚巍等[25]依据词共现特征构建特征矩阵，并进行词簇划分，实现了面向语料的主题词表自动构建；H. Liang 等[26]采用循环神经网络(RNN)技术进行文本的特征的识别与提取。③混合方法。如基于语义模板与基于统计工具相结合来提取多词表达[27]；基于同义词词林(CILIN)的特征抽取方法结合支持向量机(SVM)分类器与邻近算法分类器(KNN)完成特征提取实验[28]；基于形式概念分析(FCA)的方法获取语句中句法依赖性特征，并结合聚类方法实现文本特征的抽取[29]；等等。

对文本进行标签化处理并建立多维度标签模型，以此来计量语义内容的重要程度[30-31]，是目前量化文本内语料权重的有效方法之一。本文以文本特征提取结果为基础，将文本标签化处理应用到政府数据隐私相关文本中，来构建隐私计量模型。目前常见的文本标签化处理方法有：通过自顶向下的共现关系聚类来获取层次关系，实现内容标注[32]；基于标签分块算法[33]，挖掘文本内容间上下关系从而将文本内容标签化；基于现有的情景信息，构建多维特征的标签体系[34]，进而对模型概念进行维度划分。针对各标签的不同特点，采用不同的权重分配方案，将多维度的指标体系融合，获得综合的计量结果。如姚严志等[35]采用类信息与 TF-IDF 算法结合的方法，改进传统的文本语料权重计量方法；黄思思[36]通过提取选择文档与忽略文档的特征词，按照一定的规则进行权重调整；唐晓波等[37]基于学术论文关键词构建专长词典，提出专家专长术语的被引－逆文档权重计算方法；王佩等[38]提出基于云模型的多粒度群体决策方法，构建基于信任关系的专家权重模型；W. Li 等[39]基于模糊层次分析法，将细化指标与模糊集引入到信息安全风险的评估中。

2.3 相关研究的启示

现阶段的政府隐私保护实践中，缺乏统一的保护标准，对隐私项目进行精确计量是设立标准规范的有效途径之一。目前的隐私计量大多基于隐私主体视角，缺乏客观认知视角的补充，因此本文的研究有助于相关政策及标准的制订。

在对文本处理的过程中，为获得较高的准确率，确保所采集的隐私语义词汇具有代表性与有效性，本文采用基于文本句法结构的语言学方法进行特征提取[2,5]。对政府数据隐私进行特征提取，构建相关主题词表，完成对政府数据隐私的识别，并为政府数据隐私计量提供参考。

本文借鉴文本标签化处理的方法进行隐私计量。以构建的隐私语义词表为基础，将隐私特征词以敏感数据单元的形式呈现，构建标签层次结构计量模型[29]，采用模糊层次分析法来综合各指标[30,34]，计量不同视角下的政府敏感数据的隐私值。

3　政府数据隐私的相关文本

政府数据隐私的相关文本是指包含对政府个人数据隐私的定义、处理、论述等信息并以此为主要内容的文本。文本的视角不同，其内容也具有差异性，因此要对涉及政府数据隐私的文本进行选择。

3.1　涉及政府数据隐私的文本类型选择

本文基于客观体现隐私价值的群体视角选择文本的类型。主要有三个视角：法学理论界的视角、领域专家的视角以及隐私侵犯事件相关者的视角。从具有代表性的不同视角挖掘政府隐私边界，即是否存在敏感数据，并计算隐私值。因此要求文本具有较强的逻辑性，便于敏感数据的提取，不同视角的文本选择如下。

（1）法学理论界的视角。通过现行法律法规文本体现，包括宪法、普通法、地方法以及专门的隐私保护法等，以此为基础从权威的角度进行隐私计量。

（2）领域专家视角。通过发表在期刊中的相关论文体现，法律法规具有一定的滞后性，领域专家视角可获取最新的隐私保护内容，以此为基础从专业角度计量隐私值。

（3）隐私侵犯相关者视角。通过隐私泄露案例文本体现，导致隐私泄露案件发生的根本原因在于社会需求，需求越多，则隐私价值越大，泄露案件发生的频次也就越多，以此为基础从公众角度来计量隐私值。

3.2　相关文本的筛选

法律法规政策文本、领域专家论著文本以及隐私泄露案例文本的完整数据库都过于庞大，且文本内容重复度高。在进行敏感数据抽取与计量实验时，为保证结果的有效性，需要以代表性、时效性、相关性为主要准则，结合各类文本的不同特点对其进行筛选，并构建样本库。样本库构建方法如表 1 所示：

表 1　各类文本筛选方法

文本类型	文本筛选规则	文本筛选过程	样本库构建方法
法律法规文本	① 代表性：应考虑普遍适用性，以中央出台的法律法规政策为主； ② 时效性：所选的法律法规文本应是目前有效的，对于过时的不必收录； ③ 相关性：不仅限于隐私保护专门法，涉及个人数据隐私保护的法律法规政策都应该包括进去	在“北大法宝”中分别以“隐私”“个人信息”“个人数据”为检索词，对“中央法规”进行全文内容检索。对检索结果进行去重后获得现行有效中央法规 2 822 篇，其中法律 105 篇，行政法规 139 篇，司法解释 204 篇，部门规章 1 955 篇，党内制度 70 篇，团体及行业规定 349 篇	采取分层抽样的方式，抽出 100 篇作为本文的样本库
专家论著文本	① 时效性：应反映专家学者的最新理论成果，选择近 5 年的期刊文献； ② 代表性：所选文本应获得一定社会认可，可从核心期刊库中选取； ③ 相关性：为避免数据量过大，导致时间及成本难以接受，应根据相关度进行筛选，调整检索策略，控制检索结果的数量	在“中国知网”中进行“高级检索”，检索公式为“隐私”and“政府数据 or 政府信息”，出版年度设定为“2011－2021”，期刊来源为“北大核心”或“CSSCI”的中文期刊，共检索到结果 215 条	采取随机抽样的方式，从中抽出 100 篇作为本文的样本库
隐私泄露案例文本	① 时效性：可选择近 10 年官方权威公布的案例文本； ② 代表性：应选择能引起广泛讨论、案件信息明确、有判决结果的代表性案例； ③ 相关性：涉及到侵犯个人隐私的各类型案例都应包括在内	在“北大法宝”中以“隐私”“个人数据”“个人信息”为检索词，检索“司法案例”“案例报道”等案例文本，案例报道发布日期选择为“2011. 01. 01－2021. 11. 01”，经去重筛选后，共检索到各类隐私泄露案例文本 717 篇	采取分层抽样的方式，抽出 100 篇作为本文的样本库

经过筛选后的样本库的文本总量为 300 篇，其中法律法规政策文本 100 篇，包含：法律 4 篇，行政法规 5 篇，司法解释 8 篇，部门规章 69 篇，党内制度 2 篇，团体及行业规定 12 篇；领域专家论著文本 100 篇；隐私泄露案例文本 100 篇，按照北大法宝的案件类型分类为：典型案例 35 篇，经典案例 25 篇，普通案例 5 篇，热点案例 29 篇，司法解释性案例 6 篇。

4　面向政府数据隐私的语义词表构建

本文基于句法结构进行隐私特征提取，构建完整的政府数据隐私语义词表。词表构建采用自顶向下与自底向上相结合的方法，自顶向下是根据词表编制的目的从顶层宏观级别建立语义框架模型，提供词表顶级类别控制；自底向上则是从语料库中提取领域概念或术语，通过概念间的归类合并形成低级别的细分类别。

政府部门是政府数据的主要采集者与使用者,数据隐私类型全面且有效,因此以政府部门采集的数据隐私类型作为基础语料。基于该语料,从法律法规政策文本、专家论著文本与隐私泄露案例文本中提取相应的敏感数据信息,抽取敏感数据基础词词汇、核心动词词汇、程度词词汇、否定副词词汇等,利用自然语言处理方法与人工方法相结合的方式识别词间关系,并经过人工分类实现半自动化地构建政府数据隐私语义词表,如图1所示:

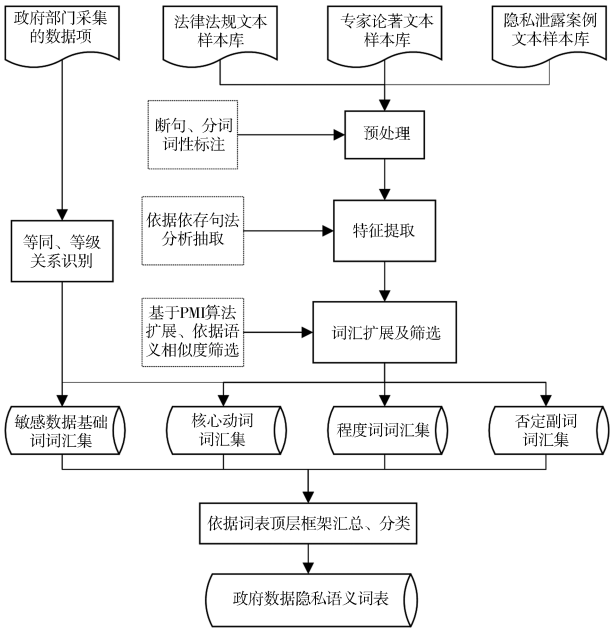

图1　词表构建流程

4.1　词表框架构建

隐私可以理解为隐私主体对隐私客体敏感性和敏感程度的认知。因此可以根据与隐私主体的关联程度将隐私客体划分为三大类别:①标识数据:隐私主体的一些标识数据,能够直接识别隐私主体的数据;②半标识数据:隐私主体的一些半标识数据,无法直接识别隐私主体,但与其他来源的半标识数据关联后可间接识别隐私主体的数据;③敏感数据:隐私主体不期望公开的关于自己的数据,主要为个人特征数据和个人行为数据。

核心动词词汇根据数据的生命周期进行分类,政府数据的生命周期一般可分为采集、保存、使用、开放4个阶段,每个阶段对数据的具体操作可扩展为若干类别,共可分为10类,分别代表了数据的产生、收集、管理、处理、保护、利用、清除、公开、验证等9个过程以及“其他”类别的核心动词。程度词包括形容词和副词,而非仅由副词组成,可以依据其修饰主体的不同而分为敏感数据的敏感程度词、核心动词程度词两类。否定词在文本中的语义表达更加强烈,如禁止、不得等,需在词表中单列出来。

4.2　隐私词汇采集方法

本文中词汇采集的流程主要有以下两种:

4.2.1　采集方法一

对不同政府部门采集的数据隐私类型进行归纳整理。在日常行政管理中,政府行政部门通常会采集个人数据。本文在选择政府部门时以涉及采集与使用个人数据的部门为主,尽量做到全面覆盖。按照政府部门的划分,不同政府部门采集数据的隐私客体不同,覆盖面广、实用价值高,具有全面性与有效性。

4.2.2　采集方法二

对涉及个人数据隐私的法律法规政策、领域专家论著以及隐私泄露案例进行文本挖掘。对文本采用逐词遍历法进行断句、分词、词性标注,基于句法结构提取文本中的敏感数据单元。将其中的动词、形容词、名词(敏感数据基础词)、副词(含否定副词)逐一分拣,并进行人工分类。

4.3　采集结果的处理及筛选

4.3.1　词间关系识别

在构建敏感数据基础词词表时,需要对收集到的敏感数据基础词汇进行整理,按照等级关系、等同关系进行识别与筛选。在构建政府数据隐私语义词表的过程中,等级关系主要有属种关系和隐私客体类别中的整部关系,等同关系的出现主要表现在不同文本中隐私项目的名称可能略有不同,但内容却基本一致,在分类的过程中可划为同一类别或者进行替代。将敏感数据基础词进行汇总整理,根据不同类属关系,归纳敏感数据单元到相应框架。

4.3.2　词汇扩展与筛选

针对核心动词、程度词以及否定副词的词汇,采用点间互信息法(PMI)进行基于文本抽取词的词汇扩展,以保障词典应用效果应增加词汇数量。经过PMI算法的扩展,导致扩展所得词汇或与原词语语义有较大差异,可利用通过词语共现关系,来确定两个词的语义关联度,采用基于《同义词词林》语义相似度的词汇筛选,获得语义相似度较高的核心动词、程度词以及否定词。

4.4　词汇采集实验

敏感数据基础词的词汇采集主要通过访问相关部门的网站,查询并统计办理各项业务所需填写或上报的个人信息项目,并结合文本中提取到的隐私项目进

行归纳与筛选。

本文共选择了 10 个相关部门，几乎涵盖了我国日常生活的所有与个人相关的政府部门，包括：①公安部门；②交通运输部门；③人事、人事组织部门；④住房与城乡建设部门、国土部门；⑤民政部门；⑥税务部门；⑦银行金融部门；⑧卫生与健康委员会；⑨教育部门；⑩邮政通讯部门。经过归纳筛选及词间关系识别后，共整理出敏感数据基础词词汇数据 178 项。结果示例如 2 所示：

表 2　敏感数据基础词词表

敏感数据基础词 隐私客体属性	敏感数据基础词 隐私数据类型	敏感数据基础词 隐私数据条目	敏感数据基础词
标识数据	标识数据	自然属性识别数据	姓名、肖像、声音、指纹……(16 项)
		社会属性识别数据	身份证号、护照号、驾驶证号……(17 项)
半标识数据	半标识数据	自然属性特征数据	性别、年龄、生日、身高……(5 项)
		社会属性特征数据	户籍、国籍、民族、出生地……(6 项)
敏感数据	家庭数据	家庭数据	婚姻信息、成长经历、家庭情况……(4 项)
	位置数据	行踪数据	出入境信息、旅客信息、定位……(5 项)
		固定位置	住址(住所)、工作单位(办公地点)(2 项)
	经济数据	财产数据	不动产信息、股票、知识产权……(20 项)
		福利数据	社会保障信息、住房公积金信息、低保信息……(8 项)
		收支数据	经济来源、消费情况、支付信息……(7 项)
	人事组织数据	人事数据	学历、学位、专业职称……(9 项)
		组织数据	纪检信息、干部信息、奖惩处分记录……(7 项)
	密码数据	密码数据	支付密码、银行账号密码、网站密码……(10 项)
	个人隐私数据	个人特征数据	行为习惯、兴趣爱好、性格特征……(4 项)
		个人内容数据	日记、遗嘱、隐私绯闻……(6 项)
		网络数据	用户注册信息、cookies、浏览记录……(8 项)
	信用与非法数据	信用数据	诚信记录、征信信息……(3 项)
		违法违规数据	刑事处罚信息、涉罪信息、不良行为史……(4 项)
	医疗数据	医疗数据	疾病病史、病历记录、精神状况……(22 项)
	通讯数据	通讯地址	联系方式、电子邮箱、IP 地址……(6 项)
		通讯内容	通话记录、手机短信、电子邮件……(9 项)

将文本中的核心动词提取并汇总，经过 PMI 扩展及筛选后，按照相应的类别进行分类归纳，构建核心动词词表。经过处理及筛选后实验收集到核心动词共 247 项。结果示例如表 3 所示：

表 3　核心动词词表

核心动词 类别	核心动词
数据产生	建立、创建、涉及、创造、产生、制作、拍摄 (7 项)
数据收集	获取、采集、收集、填写、记载、记录、载明、导出、掌握……(65 项)
数据管理	管理、存储、留存、封存、封锁、启封、归档、存档、统计……(18 项)
数据处理	修复、抢救、补救、纠错、更正、处理、加工、操作、挖掘……(30 项)
数据保护	防护、保存、保障、匿名化、去标识化、去身份化、加密……(24 项)
数据利用	应用、使用、滥用、利用、收受、购买、出售、售卖、交互……(30 项)
数据清除	删除、清除、移除、擦除、注销、毁弃、毁损、销毁、毁灭 (9 项)
数据公开	共享、整理、公开、公示、发表、发行、曝光、泄露、披露……(20 项)
数据验证	核查、审查、检查、核实、确认、认证、评估、辨认、判断……(27 项)
其他	篡改、伪造、假冒、变造、抢夺、侵犯、危害、侵害、丑化……(17 项)

程度词与否定词均为敏感数据项或核心动词的修饰词，因此否定词词表与程度词词表可一起呈现。经过处理及筛选后实验收集程度词 174 项，其中否定词 30 项，敏感数据的敏感程度词 26 项，核心动词程度词 118 项，对词汇进行情感分析评价，否定词按照最高级“5”处理，结果示例见表 4。

5　政府数据隐私计量

5.1　敏感数据单元

进行文本语料库的预处理，包括数据清洗、去除空格、特殊符号处理等，并利用中文分词技术进行分词和词性标注，去停用词；结合词性、主题相关度以及词语关联度，对触发词进行提取。在此基础上，对关键句进行提取。敏感数据单元模型如公式(1)所示：

$$DU=[V_i,U_i,(D,N)] \qquad \text{公式(1)}$$

其中，DU 为敏感数据单元；V_i 为活动词，即核心动词；U_i 为敏感数据基础词，D 表示程度修饰，N 表示否定修饰。

5.2　构建隐私计量模型

通过敏感数据单元信息，结合语义词典构建隐私

表 4　程度词及否定副词词表

程度词类别	程度词	强度值
核心动词程度词	完善、加快、严格、准确、真实、妥善、及时、如实……(118 项)	1
敏感数据的敏感程度词	低风险、弱/次之(1)、五千条以上为“情节严重”	1
	偏弱/次之(2)、弱(2)/次之、较小、较低、	2
	中风险、偏强/次之(3)、偏弱(3)/次之、一定危害、五百条以上为“情节严重”、高于(视两边敏感数据而定)	3
	强/次之(4)、偏强(4)/次之、较高、更大、更强	4
	严重危害、最敏感、高敏感、高度敏感、高风险、风险大、五十条以上为“情节严重”、强(5)/次之	5
否定词	不得、不可、不能、不被、停止、禁止、杜绝、拒绝……(30 项)	5

值计量模型。隐私值计量需要综合词频、程度值、文本力度等因素,借鉴标签化处理的过程,将各影响因素转化为计量指标,通过指标的加权融合实现计量。各指标之间引入三角模糊数的模糊层次分析法[40]进行加权处理。

基于敏感数据单元的指标包含有敏感数据项本身的指标(D_i)与语义强度指标(S_i)。

敏感数据项的词频(D_1)以及涉及敏感数据项的文本比重(D_2)可以反映隐私项在该类型文本中出现多少,出现越多则代表重视程度越高,即可能泄露或出现风险的程度越高。隐私项在语义词典中所处位置(D_3)可反映该隐私项隶属标识数据、半标识数据或敏感数据,可以直接反映出与隐私主体的关联紧密程度。

核心动词词频(S_1)可反映对敏感数据项进行处理的频次与力度,词频越高该敏感数据项的应用场景就越多,则重要程度相对就越高;程度词强度值(S_2)可直接反映文本中该隐私项因泄露产生的风险强度,强度值越高则重要程度越高

根据不同文本的类型,文本力度指标(T_i)的子指标也不相同。法律法规文本中,对文本力度指标可根据法律法规的颁布机构(T_1)及其政策力度(T_2)联合建立量化模型。在专家论著文本中,可以文章被引用次数作为论著影响力度(T_3),文章下载次数作为论著传播力度(T_4),二者结合作为文本力度评价指标。在隐私泄露案例文本中,引入案件影响力度(T_5)与发布渠道力度(T_6)来评价案例的文本力度。综合以上构成隐私计量模型,即公式(2):

$$V_i = (D_i, S_i, T_i) \qquad \text{公式(2)}$$

基于文本的政府数据隐私计量体系见图 2。

5.3　隐私计量实验

以上文构建的样本库为基础,抽取文本中敏感数据单元,进行隐私值计量实验。根据隐私计量模型,通过引入三角模糊数的模糊层次分析法将以上指标综合,计量隐私值。

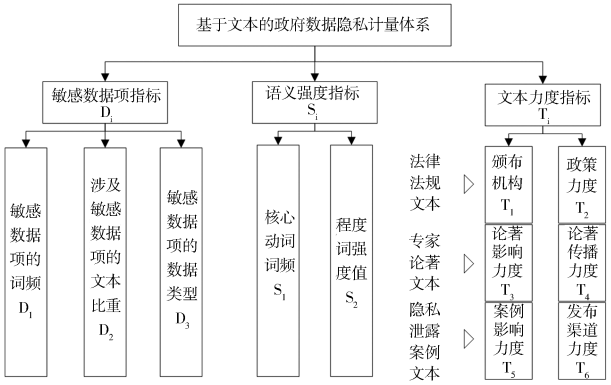

图 2　隐私计量体系框架

5.3.1　各指标的评价规则

(1)敏感数据项指标 D_i。敏感数据项指标由三个子指标组成,各子指标间按一定比例加权融合,获得该隐私项的敏感力度评级。敏感数据基础词词汇相关的指标统计如表 5 所示:

表 5　敏感数据项评价指标

敏感数据项指标 D_i	评价方法	评价结果	
D_1	词频统计	未单独提及的敏感数据项	1
		词频 10 次以内	2
		词频 11－40 次	3
		词频 41－70 次	4
		词频 71 次以上	5
D_2	文本比重	未单独提及的敏感数据项	1
		占样本总量的 20% 以内	2
		占样本总量的 20% －40%	3
		占样本总量的 40% －60%	4
		占样本总量的 60% 以上	5
D_3	分类统计	敏感数据	1
		半标识数据	3
		标识数据	5

(2)语义强度指标 Si。对敏感数据项相关的核心动词、程度词、否定副词进行统计分析。依据语义框架,计算敏感数据项对应的语义强度平均值,语义强度分为两部分,一是核心动词词频,二是程度词强度值,将两者平均值的计算结果按一定比例加权融合,获得

该隐私项的语义强度评级。核心动词由敏感数据单元中对应敏感数据项的动词词频累加得出，程度词强度值则对照程度词与否定词词表中的强度值累加得出。各子指标的计算方法如表 6 所示：

表 6 语义强度指标

语义强度指标 S_i	计算方法（超过“5”的按“5 计算”）
S_1	核心动词词频 / 敏感数据项词频
S_2	强度值 / 敏感数据项词频

（3）文本力度指标 T_i。由样本库中文本类型可知，颁布机构（T_1）包括“全国人民代表大会及常务委员会”“中共中央”“国务院”“国务院各部委”；政策力度（T_2）包括“法律”“联合文件”“行政法规”“条例”“规定”“通知”等。将两者结合构成法律法规文本力度评价指标并划分等级。将涉及隐私项的所有文本的力度值累加取均值，获得该隐私项的法律法规文本力度评级，如表 7 所示：

表 7 法律法规文本力度指标

评价指标	评价等级
国务院各部委的通知；各团体或行业的规定	1
国务院其他文件；国务院各部委部门规章；党内的制度	2
中共中央、国务院联合文件；国务院行政法规	3
最高人民检察院、最高人民法院颁布的司法解释或其他指导性文件	4
全国人民代表大会及其常务委员会颁布的法律	5

专家论著文本力度主要影响因素有两点，论著的影响力度（T_3）与传播力度（T_4），可分别用该篇文章的被引用次数以及被下载次数决定，分为 5 个等级，等级划分规则与上文敏感数据项词频统计结果评价规则一致。将涉及隐私项的所有文本力度值累加取均值，获得专家论著文本力度评价指标，如表 8 所示：

表 8 专家论著文本力度指标

论著影响力度	论著传播力度	评价等级
文章被引用频次 5 次以下	文章被下载 200 次以下	1
文章被引用频次 6 – 10 次	文章被下载 201 – 500 次	2
文章被引用频次 11 – 15 次	文章被下载 501 – 800 次	3
文章被引用频次 16 – 20 次	文章被下载 801 – 1 200 次	4
文章被引用频次 21 次以上	文章被下载 1 200 次以上	5

隐私泄露案例文本力度主要有案件影响力度（T_5）与发布渠道力度（T_6）决定。其中案件影响力度主要为案件的参照级别，包括：司法解释性案例集（与司法解释文件同等效力的案例公报）、典型案例（具有普遍指导意义或典型意义的案例公报）、经典案例（各机构以官方形式发布的案例公报）、热点案例（引起广泛关注的案例）、普通案例。案件发布渠道包括：最高人民检察院与最高人民法院、省级检察院与法院、各地方检察院与法院、权威出版机构、各级媒体。将涉及隐私项的所有案例文本的力度值累加取均值，获得该隐私项的泄露案例文本力度评级，如表 9 所示：

表 9 隐私泄露案例文本力度指标

案件影响力度	发布渠道力度	评价等级
普通案例	各级媒体	1
热点案例	司法类权威出版机构	2
经典案例	各地方检察院与法院	3
典型案例	各省级检察院与法院、其他省级部门	4
司法解释性案例	最高人民检察院、最高人民法院、其他国家级部门	5

5.3.2 指标权重计算

运用专家评价法，结合三角模糊数，对指标层和子指标层进行两两比较，构建模糊判断矩阵 A，如公式（3）与公式（4）所示：

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \quad \text{公式（3）}$$

$$a_{ij}=(l_{ij},m_{ij},u_{ij}) \quad \text{公式（4）}$$

其中，n 指每个指标中子指标的个数，a_{ij} 表示指标 i 对于指标 j 的重要程度，以三角模糊的形式（例如公式（3））表示，依据三角模糊数的结构，l 表示下界值，m 表示中值，u 表示上界值。参考已有的标度法[41]构建模糊判断矩阵，重要性标度方法如表 10 所示：

表 10 重要性标度方法

含义	三角模糊数
非常重要	（2,5/2,3）
特别重要	（3/2,2,5/2）
相对重要	（1,3/2,2）
一般重要	（1/2,1,3/2）
相同	（1,1,1）
指标 i 与 j 比较的判断为 a_{ij}，指标 j 与 i 比较的判断为 $a_{ji}=1/a_{ij}$	倒数

多专家评判结果取三角模糊数平均值作为综合三角模糊数，构建判断矩阵 A，基于模糊数比较原理，通过三角模糊数的模糊层次分析算法[42]，可计算各指标权重，如表 11 所示：

表 11 指标层判断矩阵及权重

指标	Di	Si	Ti	权重 W
Di	（1,1,1）	（1.67,2.17,2.67）	（0.83,1.33,1.83）	0.49
Si	（0.37,0.46,0.61）	（1,1,1）	（0.61,1,1.83）	0.23
Ti	（0.56,0.78,1.33）	（0.56,1,1.67）	（1,1,1）	0.28

同理可求得子指标层各指标权重。敏感数据项指标中各子指标权重分别为:{0.4,0.27,0.33};语义强度指标中各子指标权重分别为:{0.58,0.42}。按照此比例将各项指标的值加权融合,获得具体的隐私项文本隐私值。

5.4 隐私计量结果及分析

5.4.1 计量结果简述

计量实验针对三个文本库分别进行,每个隐私项可能出现三个隐私计量模型,代表不同认知视角,即法学理论界、领域专家、隐私侵犯者等对敏感数据敏感程度的认知。部分计量结果(取前5名)如表12所示:

表12 基于文本的隐私值计量结果(部分)

	法律法规文本	隐私项(隐私值)专家论著文本	隐私泄露案例文本	排名
标识数据与半标识数据	身份信息(3.41)	身份证号(3.34)	姓名(3.52)	1
	身份证号(3.21)	身份信息(3.30)	身份证号(3.13)	2
	姓名(3.17)	姓名(3.24)	身份信息(3.02)	3
	人脸(2.96)	照片(2.95)	生物识别数据(2.84)	4
	照片(2.95)	生物识别数据(2.87)	基因数据(2.80)	5
敏感数据	网络数据(2.91)	通信数据(2.65)	支付密码(3.60)	1
	通话记录(2.86)	住址(2.60)	网络发布数据(2.97)	2
	通信数据(2.82)	福利数据(2.55)	通信数据(2.61)	3
	经济数据(2.67)	财产数据(2.54)	个人信用数据(2.59)	4
	网络发布数据(2.62)	网络发布数据(2.48)	征信数据(2.49)	5

由表12可知,在标识数据与半标识数据中,三种文本的计量结果没有明显差别,“身份信息”“身份证号”“生物识别数据”“姓名”在三个文本中均占据前列,法律法规与专家论著文本中的“照片”数据以及隐私泄露案例文本中的“基因数据”也均为“生物识别数据”的一部分。无论从法律与学者的角度,还是隐私泄露情况的角度,以身份证为主的身份数据和个人生物识别数据都是隐私保护问题的重点及难点。

三种文本计量结果的差别主要体现在基础敏感数据上:

(1)基于法律法规的角度,保护公民的个人隐私与个人财产不受侵犯是法律的职责所在。保护的重点主要体现在:①个人内容隐私,如公民在网络活动中涉及的数据、私人通信数据、网络发布数据等;②有关个人财产的经济数据。在相关法律法规中,包含“病历”数据在内的“个人医疗健康数据”隐私值也相对较高,通常与“个人财产数据”的重要程度相当。

(2)专家视角下,关注的隐私问题在法律的基础上更加具体化,如与个人生活财产关系密切的“住址”数据与“福利数据”等;其次针对敏感数据项的讨论也更加分散化,各类型的隐私在文本中均有呈现,导致计量结果的分差不大。

(3)隐私侵犯相关者的角度则更能体现隐私的风险性。以“支付密码”为主的个人支付数据与密码数据成为了重中之重,隐私值远远超过第二名的“网络平台发布数据”,前者直接威胁到个人的财产安全,后者则因大数据与人工智能产业的兴起造成了网络用户的个人数据大量违规滥用。此外“个人信用数据”也在隐私泄露案例中频繁出现,信用数据被滥用冒用也会给个人生活带来极大的不便,应当引起重视。

5.4.2 结果分析与讨论

从三种文本计量结果的共性来看,按照敏感数据基础词词表的数据类型将相同类型的敏感数据项目累加后,结果的相似度进一步提高。以“通信地址”数据与“通信内容”数据为主的“通讯数据”以及包含有“财产数据”“收支数据”与“福利数据”的“经济数据”,在出现的频次以及文本占比上都大幅领先,与最终的计量结果相差不大。包含了“网络平台发布数据”的“个人内容数据”在文本中所占比例并不突出,但由于其程度值与文本力度相对较高,在最终计量结果中排名靠前,这也与以往隐私价值计量研究中网络用户将“个人内容数据”价值视为最高[16]的结果相吻合。

从三种文本计量结果的差异性来看,法律法规文本与专家论著文本的计量结果更为接近,隐私泄露案例文本则稍有不同。包括“病历”数据在内的“个人医疗健康数据”在法律法规文本与专家论著文本中多次出现,程度修饰词的强度值也相对较高,此类数据(尤其有关传染性疾病、性器官疾病、性功能障碍性疾病等特殊疾病信息)的泄露往往会对个人的生活产生较大影响[8],因此此类数据隐私值排在前列,仅次于表12中前5的敏感数据。但在实际隐私泄露案例中,关于隐私的犯罪往往是以金钱为目的,伴随诈骗、盗窃、非法交易等经济犯罪,因此关乎个人财产与个人联系方式的隐私数据隐私值更高。

基于文本的隐私计量最主要的问题是具体的隐私项与隐私分类不够明确和完整。在文本中,通常以“隐

私”“个人信息”或“个人数据”的总称进行叙述。以“隐私”为主题的相关论文大部分研究对象是隐私保护政策及措施,缺乏对隐私内容的直接探讨。在计量结果中,“身份信息”在三个文本中均排名前列,针对“身份信息”具体包含的数据项却没有完整的解释。“通信数据”也是极为重要的隐私项之一,“通信数据”中包含了“通信地址数据”以及“通信内容数据”两个大的类别,其中更有众多具体的隐私项,两个类别的重要性有何区别也很难具体计量。但整体看来,法律、专家及公众不同视角下隐私关切的差异也反映出隐私保护措施的滞后性。基于文本的隐私计量具有客观的参考价值,与基于隐私主体认知的隐私价值计量互为补充,为精确计量隐私提供了基础。

6 结语

本文主要探讨了基于文本的政府数据隐私特征识别及计量。在敏感数据识别方面:一是涉及范围广,涵盖了法律法规、专家论著、隐私泄露案例等全视角的相关文本,敏感数据相关词汇收集较为全面;二是敏感数据单元构建相对完整,包含了核心动词、敏感数据基础词、程度词、否定副词等。在隐私计量方面:本文创新性地以非常规的隐私计量手段,从相关文本入手,计量客观视角下的隐私值,计量过程不受人为主观因素限制、相对客观,与基于隐私主体认知的隐私计量形成了对照参考,为规范政府数据隐私保护标准提供了新的研究视角。由于文本描述的特点,导致隐私类别划分不够明确;加之统计方法不够完善,各隐私项单独统计,没有按照类别进行合计。但从结果来看,基于文本的隐私计量结果准确地反映了三个视角下的隐私关切。在未来的研究中应积极改进,使研究更加完善。

参考文献:

[1] 黄如花,吴子晗. 中国政府数据开放共享政策的计量分析[J]. 情报资料工作,2017(5):6-12.
[2] 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.
[3] 赵金旭,郑跃平. 中国电子政务隐私保护问题探究——基于70个大中城市政府网站的“隐私声明”调查[J]. 电子政务,2016(7):81-93.
[4] 商希雪,韩海庭. 政府数据开放中个人信息保护路径研究[J]. 电子政务,2021(6):113-124.
[5] 丁红发,孟秋晴,王祥,等. 面向数据生命周期的政府数据开放的数据安全与隐私保护对策分析[J]. 情报杂志,2019,38(7):151-159.
[6] 张聪丛,郜颖颖,赵畅,等. 开放政府数据共享与使用中的隐私保护问题研究——基于开放政府数据生命周期理论[J]. 电子政务,2018(9):24-36.
[7] KIM J, NAM C, KIM S. The economic value of personal information and policy implication[C]// Proceedings of the 26th European regional ITS conference. Los Angeles: ITS Press,2015.
[8] 臧国全,贾瑞莹. 医疗数据中病种隐私的计量与分析[J]. 现代情报,2020,40(5):161-168.
[9] HANN I H, KAI L H, LEE T, et al. Online information privacy: measuring the cost-benefit trade-off[EB/OL]. [2021-12-30]. https://aisel.aisnet.org/icis2002/1.
[10] EGELMAN S, FELT A P, WAGNER D. Choice architecture and smartphone privacy: there's a price for that[C]// BÖHME R. The economics of information security and privacy. Heidelberg: Springer,2013:211-236.
[11] 邓胜利,赵海平. 信息泄露情境下的个人信息价值评估及个体差异:基于离散选择模型的实证研究[J]. 情报学报,2019,38(3):266-276.
[12] 黄逸珺,陆桐,闫强. 电子商务网站个人信息价值评估[J]. 北京邮电大学学报(社会科学版),2017,19(5):33-41.
[13] BENNDORF V, NORMANN H T. The willingness to sell personal data[J]. The scandinavian journal of economics, 2018, 120(4): 1260-1278.
[14] 张凯亮,臧国全. 泄露概率情境下的个人数据隐私计量研究[J]. 图书情报工作,2021,65(9):62-69.
[15] SPIEKERMANN S, BAUERC, KORUNOVSKA J. Psychology of ownership and asset defense: why people value their personal information beyond privacy[J]. Social science electronic publishing, 2012, 4(1):41-47.
[16] 臧国全,张凯亮,闫励. 个人数据价值计量研究——基于改造的BDM机制[J]. 图书情报工作,2020,64(7):103-109.
[17] 邵卫,化柏林. 基于依存句法分析的科技政策领域主题词表无监督构建[J]. 情报工程,2020,6(6):33-44.
[18] 郑新曼,董瑜. 基于科技政策文本的程度词典构建研究[J]. 数据分析与知识发现,2021,5(10):81-93.
[19] LI X F, ZHAO L L, WU L H. A feature extraction method using base phrase and keyword in Chinese text[C]//Proceedings of the 2008 3rd international conference on intelligent system and knowledge engineering. Xiamen: IEEE,2008:696-700.
[20] 唐晓波,肖璐. 基于依存句法网络的文本特征提取研究[J]. 现代图书情报技术,2014(11):31-37.
[21] 葛斌,封孝生,谭文堂,等. 基于多层最大熵模型的句子主干分析[J]. 计算机科学,2010,37(12):156-160.
[22] 涂安龙. 一种CM-RS文本特征提取方法研究[D]. 武汉:华中师范大学,2012.
[23] 曹钲晨. 基于海量数据分析的汉语句式特征提取及应用[D]. 济南:山东大学,2021.
[24] DIETZ E A, VANDIC D, FRASINCAR F. Taxolearn: a semantic approach to domain taxonomy learning[C]// Proceedings of the 2012 IEEE/WIC/ACM international joint conferences on Web intelligence and intelligent agent technology. Washington DC: IEEE Computer Society, 2012:58-65.
[25] 安亚巍,操晓春,罗顺. 面向语料的领域主题词表构建算法[J]. 计算机科学,2018,45(S1):396-397,410.

[26] LIANG H, SUN X, SUN Y, et al. Text feature extraction based on deep learning: a review[J]. Eurasip journal on wireless communications & networking, 2017, 211:1－12.
[27] 肖健,徐建,徐晓兰,等.英中可比语料库中多词表达自动提取与对齐[J].计算机工程与应用,2010,46(31):130－134,187.
[28] LI X F, ZHAO L L. A multilayer method of text feature extraction based on CILIN[C]// Proceedings of the 2008 international conference on computer science & information technology. Singapore: IEEE, 2008:48－52.
[29] CIMIANO P, HOTHO A, STAAB S. Learning concept hierarchies from text corpora using formal concept analysis[J]. Journal of artificial intelligence research, 2005, 24(1): 305－339.
[30] 毕崇武,叶光辉,李明倩,等.基于标签语义挖掘的城市画像感知研究[J].数据分析与知识发现,2019,3(12):41－51.
[31] 夏立新,曾杰妍,毕崇武,等.基于LDA主题模型的用户兴趣层级演化研究[J].数据分析与知识发现,2019,3(7):1－13.
[32] HEYMANN P, GARCIA-MOLINA H. Collaborative creation of communal hierarchical taxonomies in social tagging systems[R]. Palo Alto: Stanford InfoLab Publication Server, 2006.
[33] 刘苏祺,白光伟,沈航.基于用户自描述标签的层次分类体系构建方法[J].计算机科学,2016,43(7):224－229,239.
[34] 阮雪灵. 基于用户画像的普适推荐方法与模型研究[D].武汉:武汉纺织大学,2021.
[35] 姚严志,李建良.基于类信息的TF-IDF权重分析与改进[J].计算机系统应用,2021,30(9):237－241.
[36] 黄思思.基于特征词权重变更的检索优化策略[J].情报科学,2016,34(7):70－75.
[37] 唐晓波,周禾深,李诗轩,等.基于被引－逆文档权重的专家专长识别与分析——以图情领域为例[J].图书情报工作,2021,65(15):111－119.
[38] 王佩,张婧,张威威.基于云模型和多层权重求解的多粒度语言大群体决策方法[J].控制与决策,2021,36(9):2257－2266.
[39] LI W, LIANG Y, WANG W, et al. Research on security risk assessment based on the improved FAHP[C]// Proceedings of the 3rd annual international conference on cloud technology and communication engineering. Wuhan: IOP Publishing, 2020.
[40] 蒋斌,梁小安,高杨军,等.基于可靠度确定属性权重的三角模糊数多属性决策方法[J].模糊系统与数学,2021,35(4):113－123.
[41] KAHRAMAN C, ERTAY T, BUYUKZKAN G. A fuzzy optimization model for QFD planning process using analytic network approach[J]. European journal of operational research, 2006, 171(2):390－411.
[42] CHANG D Y. Applications of the extent analysis method on fuzzy AHP[J]. European journal of operational research, 1996, 95(3):649－655.

作者贡献说明：

臧国全：确定选题,提出研究思路,论文审核与修订；

王家振：论文写作与修改,数据收集与整理,实验设计与实现；

毕崇武：论文框架设计,论文修订；

耿瑞利：研究方法与理论设计。

Research on Sensitive Data Identification and Privacy Measurement in Government Data

Zang Guoquan[1,2] Wang Jiazhen[1] Bi Chongwu[1,2] Geng Ruili[1,2]

[1] School of Information Management, Zhengzhou University, Zhengzhou 45001

[2] Research Institute of Data Science, Zhengzhou City, Zhengzhou 450001

Abstract: [**Purpose/Significance**] Through the analysis of government data privacy related texts, designing sensitive data identification scheme, building a privacy measurement model, and measuring the privacy value of sensitive data, this paper provides a theoretical basis for government data privacy protection. [**Method/Process**] First, filtered the relevant text of government data privacy to build a sample library; Then, according to the syntactic structure of the text, words such as sensitive data items, core verbs, degree words, negative words were extracted, it constructed the semantic vocabulary of government data privacy; Finally, based on the sensitive data unit composed of the above words, it constructed privacy measurement model. [**Result/Conclusion**] This method is based on privacy related texts, accurately extracts the sensitive data of government data, objectively measures the privacy value of government data objects, and provides support for the privacy risk prevention and privacy protection standardization of government data.

Keywords: government data　data privacy　personal privacy　semantic vocabulary　privacy measurement