第 32 卷 第 4 期 石家庄铁道大学学报(自然科学版) Vol. 32 No. 4
2019 年 12 月 Journal of Shijiazhuang Tiedao University(Natural Science Edition) Dec. 2019

跨部门政府数据共享中的隐私保护研究

史雅涓[1]， 朴春慧[1]， 颜嘉麒[2]

(1. 石家庄铁道大学 信息科学与技术学院，河北 石家庄 050043；
2. 南京大学 信息管理学院，江苏 南京 210023)

摘要：随着大数据时代的到来，政府部门越来越重视运用技术手段深度挖掘政府数据资源的价值。与此同时，政府部门对于政府数据共享的需求也在不断提高。然而，目前政府数据共享中存在着不同程度的隐私泄露风险。因此，如何在保护公民隐私的前提下共享政府数据成为提高政府治理能力和服务水平的一个挑战性问题。从政府部门间数据共享隐私保护需求出发，根据政府数据类型多样、属性复杂等特性，提出了基于中心点聚类的改进 K 匿名数据共享方法(KMedoid-based KADS)。首先，为减少非需求属性数据的共享范围，对拟共享原始数据表进行预处理，基于属性相关度将其划分为多个数据表；然后，利用基于中心点的记录相似度聚类算法处理划分得到的数据表，生成初步满足 K 匿名要求的聚类结果表；最后，利用 K 匿名簇共享算法，根据数据资源请求部门的共享要求，为其提供需求数据所在的 K 匿名数据表。通过与经典 K 匿名 Incognito 算法进行实验比较，表明提出的 KMedoid-based KADS 算法能有效减少信息损失量，提高共享数据可用性。

关键词：政府数据共享；隐私保护；雾计算；K-匿名；聚类

中图分类号：TP391 **文献标志码**：A **文章编号**：2095-0373(2019)04-0116-09

0 引言

大数据时代的来临意味着数据同物质及能源一样，成为人类社会发展所必需的战略性资源。作为公共服务的主要提供者，政府有责任向公众公开政府信息，保障公众的知情权；同时，作为公共数据资源的最大所有者，政府也有义务向社会开放数据资源，充分利用政府数据的巨大优势。不仅如此，政府在维持社会正常运行、保证个体生存发展的前提下，也应给予相关组织或部门分享部分政府数据的权利。当下各级政府部门都在积极部署政务云平台，以推动政务信息系统整合共享建设进程[1]。政务云环境下的公民隐私信息保护问题也成为政府部门和公众难以忽视的焦点问题[2]。在政府数据共享过程中，就存在着隐私泄露的可能。从政策角度分析，目前我国现行政策中有对个人隐私保护的说明，但是缺乏系统性和可执行性[3]。从共享方式角度分析，难以找到绝对可信的数据资源共享平台，以作为政府部门间数据共享的媒介。从技术角度分析，用于数据共享的匿名隐私保护技术存在着大量的信息损失，造成数据可用性差，直接影响了部门之间数据的使用。因此，对于政府机构而言，迫切需要找到合适的方法以解决数据发布共享过程中存在的隐私泄露问题。

目前，针对政府数据发布共享情境下的隐私保护研究相对较少。针对政府数据共享发布过程中出现的隐私泄露风险，大多数研究都集中于分析现有的隐私风险、完善相应法律制度。对于政府数据发布共享场景下的具体应用框架及适用的隐私保护方法关注度不足。因此，对于政府机构而言，迫切需要找到合适的方法以解决数据共享过程中存在的隐私泄露问题。

收稿日期：2018-05-10 **网络出版日期**：- **责任编辑**：翟晓玲 DOI：10.13319/j.cnki.sjztddxxbzrb.20180510002
网络出版地址：http://kns.cnki.net/kcms/detail/13.1402.N.20191120.1415.020.html
作者简介：史雅涓(1993—)，女，硕士研究生，主要从事计算机软件及计算机应用研究。E-mail：syjsghhz@sina.com
史雅涓，朴春慧，颜嘉麒.跨部门政府数据共享中的隐私保护研究[J].石家庄铁道大学学报：自然科学版，2019，32(4)：116-124.

1　政府数据共享中的隐私保护研究

1.1　政府数据共享中的隐私问题研究

政府作为政务数据的采集者、政务信息的拥有者以及政务工作的管理者,掌握了大量有价值的数据,在大数据浪潮中具有天然的信息优势。在我国,随着电子政务的普及应用,基于政府门户网站的数据发布已广为人知,但政府数据共享尚处于起步初期,很多概念及策略仍处在探索阶段。政府数据共享是指在行政单位范围内,对内共享不适合向社会开放的、有价值的数据。政府数据共享增强了政府部门间数据融合,提升了原有数据价值,从而达到提高政府工作质量的目的。

与英国等发达国家相比,我国政府数据共享尚处于初级阶段,还处在较低水平,数据共享过程仍存在着诸多问题。为促进政府数据开放共享发展,解决开放共享过程中出现的各种问题,我国政府出台了一系列相关法律法规来保证数据的安全性与隐私性。例如,《中华人民共和国网络安全法》于 2016 年 11 月由全国人大常委会审议通过,该法律的颁布,表明出国家对于数据安全与隐私保护的重视程度,有助于推动政府数据资源开放共享进程、加快技术创新步伐、提高经济发展水平[4];2017 年 5 月,国务院印发了《政务信息系统整合共享实施方案》,其中明确提出要尽快完善个人隐私信息保护的法律法规,保障政务资源使用中的个人隐私[5]。

1.2　政府数据共享中的隐私保护

由于政府数据种类繁多,针对不同的应用场景,数据共享的程度、发挥的作用也不尽相同。政府数据通常包含大量的个人相关数据,例如,个人的收入、纳税、房产、信用、违法违规记录等,这些信息均可被视为个人敏感信息。若这些敏感信息发生泄露,会给个人、企业甚至国家带来巨大的安全威胁。因此,本文研究政府数据共享中的公民隐私保护技术解决方案,探寻合理的数据共享方法,以此减轻了公民隐私顾虑,更好地促进政府数据共享的发展。针对政府数据共享中存在的隐私泄露风险,研究人员从法律、管理、信息技术等层面上都进行了探讨分析。张晓娟等[6]分析了美国政府在隐私保护立法方面的措施,为进一步完善我国个人隐私保护相关法律体系提供了思路。刘凌等[7]针对政府数据开放过程中存在的隐私泄露问题,通过总结分析国内外研究现状,提出了一种先导性的政府数据隐私保护分析框架。王芳等[8]分析了政府数据跨部门共享中出现的问题及其产生原因,从政策角度提出了隐私保护问题的解决方案。黄如花等[9]使用内容分析法分析了国家政策文件,发现我国政府数据开放共享政策尚未形成完整体系,仍需进一步完善个人隐私保护政策。

虽然上述方法进一步明确了政府数据共享发布中存在着隐私泄露风险的事实。但是,并未根据政府数据开放共享中的具体应用场景,从技术角度给出可靠的实施方案,单纯的背景分析及政策指向难以打消公民的隐私顾虑。

1.3　常用的隐私保护数据技术

隐私保护技术是指利用技术手段处理原始数据集,确保攻击者即便获取到相关背景知识,依旧不能识别出目标个体的敏感信息。在数据发布共享中,因为难以预估攻击者所了解的背景知识多少,所以并没有绝对安全的隐私保护技术。在数据发布共享中,研究应用隐私保护方法的主要目的是:在不泄露隐私的基础上,尽可能地提高数据的准确性。

常用的隐私保护技术可以分为以下 3 类[10]。

(1)基于数据加密的技术。通常是利用加密技术对个体敏感属性进行相应处理[11-12]。由于政府信息发布或数据共享的原则是尽可能地提供原始数据,这种方式一般并不适合于政府数据发布共享场景。

(2)基于数据失真的技术。通常采用扰动失真的方式对敏感属性进行处理,例如添加噪声[13-14]、交换[15]、随机化[16]等。这种方式在一定程度上保持了数据的统计特性,经常被应用于统计领域。对于政府数据发布场景,该方式能够满足公众对政府数据知情权的需要。

(3)基于限制的发布技术。通常是采用降低发布精度的方式处理敏感数据集,例如数据泛化[17]、隐匿[18]、聚类[19]等。该类技术主要集中于研究数据匿名化方法。对于政府数据共享场景,该方式保证了数

据的原有特性,更适合于政府部门间使用。

分析政府数据共享中存在的隐私泄露风险,构建基于政务云 + 雾计算的政府数据共享隐私保护框架,提出了基于中心点聚类的 K 匿名数据共享方法,进一步减轻了公民隐私顾虑,提高了共享数据质量,进而促进了政府数据共享更快更好地发展。

2 政务云十雾模式的隐私保护政府数据共享框架

2.1 政府数据共享的政务云十雾计算模式

政务云 + 雾计算模式,简单来说,就是充分利用政务云与雾计算优势,整合两者资源,相互协作共同推进政府数据公开,提升政府治理水平。政务云 + 雾计算模式主要分成 3 个部分:

(1)用户层是由第三方社会组织、企业、公众所拥有的智能终端和这些智能终端产生的海量数据所构成。一般来说,认为政府机关采集数据是可信的,数据采集存储过程不存在隐私泄露现象。

(2)政府雾层是由政府目前拥有的大量软硬件设备所组成。政府各部门通过其自身设备优势,建立本部门政府雾服务器,管理其自身业务,并向政务云层提供资源整合共享的支持。政府雾层提供数据存储、计算、隐私保护的功能,在雾层不仅可以对接收到的用户层数据进行本地实时分析处理,而且可以在部门内部直接进行数据隐私处理,降低隐私数据泄露风险。

(3)政务云层接收来自雾层隐私数据库的数据,并进行整体性分析处理。政务云层既可以满足政府机构对于高性能计算的要求,又能促进公共服务水平。例如在政务云层,政府部门可以利用云的高性能服务器分析处理大量政府数据,辅助人工决策;可以将数据进行发布公开,弥补之前的公众与政府间的数据不对称现象;可以提供多种应用服务,简化传统办事流程、方便解决民生问题。

对于政府机构来说,使用政务云 + 雾计算模式的优势在于:充分利用政府内部现有设备资源,减少大量购买高性能设备的支出;充分利用相互协作的多方终端或者边缘设备共同执行计算或存储功能,减少政务云对于高带宽、高性能的需求,提高了数据处理效率,降低政务云建设成本;充分利用雾计算提供的隐私保护功能,提高政府数据发布共享的安全性与隐私性,有助于增强政府公信力。

因此,利用政务云 + 雾计算模式有助于政府解决信息化建设过程中存在的诸多问题。例如,自媒体时代的发展,促使互联网已成为了网民发表意见的集散地,为保障国家安全、维护政府形象、促进民生改善,政府必须及时了解舆情发展动态以判断未来发展趋势。此时,实时性就成为了政府舆情监测的关键。通过使用政务云 + 雾计算模式,利用雾计算在政府本地就可以对舆情信息进行过滤、汇总,并对事件进行实时的处理。后续再将统计数据传输给政务云端,借助云平台强大的计算能力对多方信息进行大数据分析,帮助预测舆情发展趋势,指导政府下一步决策。

2.2 政务云十雾模式的隐私保护数据共享框架

基于现有的政务云平台基础架构,结合雾计算的思想以及隐私保护的需求,设计了政务云 + 雾模式的隐私保护政府数据共享框架,如图 1 所示。政府数据共享层面上的隐私服务只涉及政府雾层部分。由于政府部门共享的数据多为不宜向社会开放的、有价值的、高敏感性的数据,当其它部门未妥善利用该数据时,可能造成隐私信息的泄露。为防止这种现象的发生,也需要在政府雾层对共享数据进行隐私化处理。当数据隐私保护后,政府部门就可以通过雾雾链接直接进行数据共享。

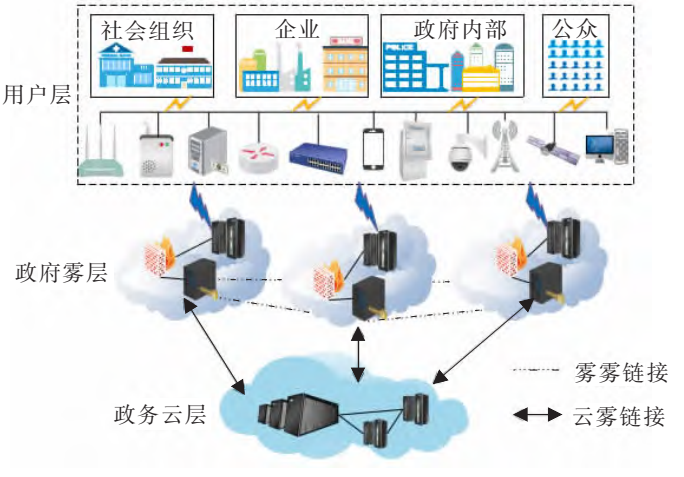

图 1 政务云十雾模式的隐私保护政府数据发布共享框架

2.3　基于雾层的隐私保护政府数据共享流程

利用雾雾链接不单可以使政府方便地处理本部门事务,还能将各部门联系在一起,帮助提高业务协同处理能力及办事效率。雾计算的使用改变了原有的政府数据共享方式,新的方式不再基于数据资源共享平台,而是利用政府雾链接关系直接共享数据。这一共享方式,需要政府着重考虑隐私保护问题。

原有的政府数据共享方式在面对敏感信息保护方面,主要依托一个前提假设,即数据资源共享发布平台能够保证不会泄露或窃取用户信息。然而,无论是从政策还是技术角度分析,都没有绝对安全、隐私的数据资源共享发布平台。借助于政务雾服务,政府部门不再需要数据资源共享平台作为媒介,部门之间可以利用雾雾链接快速共享传播各类信息事务。与此同时隐私保护技术的加入避免了公民隐私信息泄露,保证了关键信息传递的及时性和隐私性。从长远来看,政府部门间的雾雾链接共享方式需要寻求可靠、安全的隐私保护方法及实现方式,以达到隐私保护数据共享的目的。图 2 为基于雾层的隐私保护政府数据共享流程。

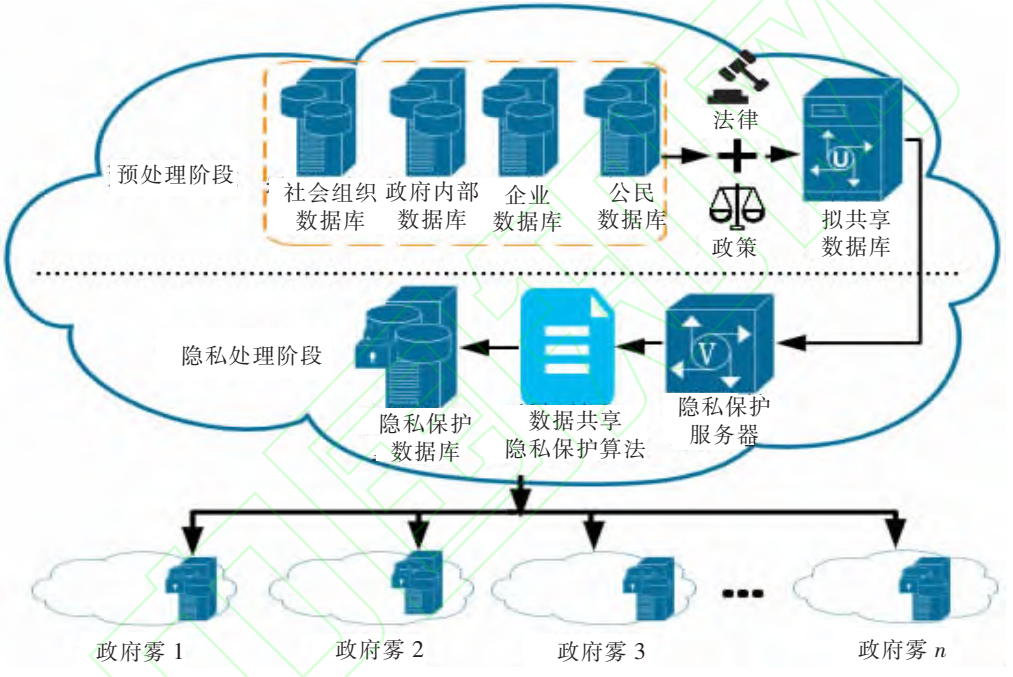

图 2　基于雾层的隐私保护政府数据共享流程

具体来说,基于雾层的隐私保护政府数据共享流程包括:

(1)政府通过对其数据库查询,挑选符合政策和法律规范的拟发布数据;

(2)利用隐私保护服务器执行数据共享隐私保护算法,并将得到共享数据存于隐私保护数据库中;

(3)依照申请情况,利用雾雾链接将数据共享给其它政府部门。

通过分析基于雾层的隐私保护政府数据共享流程,有助于确立政府数据共享机制。假设政府数据共享过程涉及 2 个实体,数据资源提供者与数据资源需求者都为政府行政部门。数据资源需求者要向数据资源提供者申请所需共享数据,在审核通过后,数据资源提供者就可在其隐私数据库查找所需数据表,并将所查数据涵盖的数据表统一共享给数据资源需求者。如图 3 为政府数据共享机制。

3　基于中心点聚类的 K 匿名数据共享方法(KMedoid-based KADS)

由于政府数据种类较多,每个种类的作用也不尽相同。相同的数据应用于不同的场景中,在含义上也可能发生变化。针对政府数据的特殊性,将政府数据集属性定为 3 类:(1)标识符属性,例如纳税人姓名、身份证号、银行卡号码等能直接识别出特定个人的属性集合。(2)敏感属性,例如企业年收入、购房人购房支出、纳税人纳税金额等个人不愿对外披露的敏感属性集合。(3)非敏感属性,上述属性之外的所有

属性集合。

基于中心点聚类的K匿名数据共享方法主要分为3个步骤。

Step 1:数据属性预处理。利用中心点聚类算法将敏感属性与非敏感属性按照相关度大小划分开,使得相关度较大的非敏感属性划分为一组,并且选取较少数量的非敏感属性与敏感属性划分为一组。数据属性预处理在降低数据维度提高算法执行效率的同时,可以有效平衡数据隐私性与可用性之间的关系,为后续数据共享算法奠定基础。

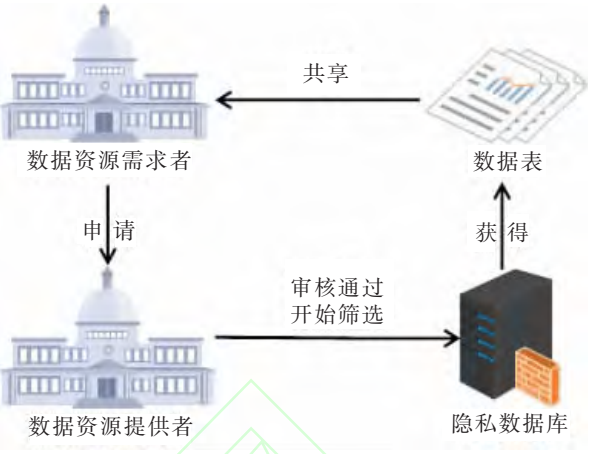

图3 政府数据共享机制

Step 2:基于中心点的记录相似度聚类算法。通过使用聚类技术,利用Gower's系数计算记录间的距离,可将数据记录划分成若干个簇,并使得同簇记录间具有较高的相似度、异簇记录间具有较高的差异性。由于同簇记录相似程度高,记录之间对应属性值更为接近,后续对各簇进行K匿名处理时,可以降低属性泛化程度,保证数据的可用性。

Step 3:K匿名簇共享算法。使用泛化技术对未完全符合K匿名要求的簇进行处理,得到完全K匿名簇,并将匿名数据以数据表的形式存于隐私数据库中。根据政府其它部门数据共享需求,在本部门隐私数据库中筛选所需属性列对应的数据表,并将所查属性覆盖的所有数据表统一共享给需求部门。

3.1 数据属性预处理

数据属性预处理需要考虑两方面的问题,一是如何对属性列进行划分,二是如何选取合适的划分数量。

3.1.1 数据属性列划分

选取K中心点聚类算法作为数据属性列划分方法。K中心点聚类算法的实现很大程度上取决于距离的计算。若使用K中心点聚类算法对政府数据属性进行划分,就需要找到合适的指标以衡量数据属性之间的距离。借助克莱姆相关系数来衡量2个属性间相关度的大小,并根据相关度与距离的关系,得到属性间距离。克莱姆相关系数被设计衡量2个分类型属性相关性的强弱,取值范围为[0,1][20]。克莱姆系数越靠近1,代表2个分类属性的相关性越强,越靠近0相关性越弱。其中克莱姆相关系数的计算公式为

$$V=\sqrt{\frac{x^2}{n\times\min[(R-1),(C-1)]}} \tag{1}$$

式中,R和C分别代表2个分类型属性的枚举数值;n为数据量;x^2为pearson卡方统计量。因此,结合数据隐私性和可用性的双重需求,本章利用K中心点聚类算法实现数据属性列划分的具体流程如下。

Step 1:计算各属性列间相关度的大小。若政府数据中存在连续型属性,则需要预先对该属性进行离散化处理,即将连续型属性按照属性值大小划分为多个大小相等的区间段。此时,每个区间段都可看作分类型属性的取值,继续利用克莱姆相关系数进行计算即可。

Step 2:找到与敏感属性列相关度最高的属性列,并将两者合并为一个集合。

Step 3:利用K中心点聚类算法对剩余属性列进行划分。其中,距离的计算方法为$D=1-V$,即克莱姆相关系数越小,距离越大,越难划分至一个集合中。

3.1.2 选取合适的划分数量

在使用K中心点算法进行聚类时,需要事先给出聚类个数,即K值。由于属性划分数直接影响着后续操作的顺利进行,找到相应的评判指标,确定划分数量是十分必要的。本节通过选用轮廓系数值判断聚类效果,帮助确定划分数量。轮廓系数可对任意距离进行度量(如欧式距离、马氏距离等),对于数据属性列划分所采用的距离计算方式也同样适用。轮廓系数通过分析内聚度和分离度这2个因素来实现聚类效果的判断[21]。利用轮廓系数,有助于找到聚类数目下最优聚类结果。在实际应用中,为尽可能地保护数据的原始性,k值的取值一般也不会很大,通过简单判断就可得到最优k值。为避免出现局部最优解

的情况,所有出现的值都需要多次运行 K 中心点算法,并且计算每一个 k 值对应的平均轮廓系数,最后选择拥有最大轮廓系数的 k 值作为最终划分数。

因此,通过数据属性预处理,将原始数据表按照最优属性划分的方法进行了划分,得到了多张属性相关度最高的数据表。

3.2　基于中心点的记录相似度聚类算法

根据上一步操作得到了多张属性关联度最高的数据表,以其中一张数据表为例,介绍如何对表中记录进行聚类以满足 K 匿名的初步要求。其余多张数据表均可依照下述方法进行处理。

依然使用中心点聚类算法解决政府数据记录相似度聚类问题。基于中心点的记录相似度聚类算法核心思想是:将包含 k 条记录的数据表,利用聚类技术聚成多个簇,使得每个簇中的记录数目至少为 k 条,且要求各簇内部记录差距最小。因此,基于中心点的记录相似度聚类算法实现过程主要分为以下 2 个步骤:

Step 1:结合有效的记录距离度量方式,利用中心点聚类算法产生各簇;

Step 2:判断簇中的记录数目,对并不满足数目要求的簇进行调整。

3.2.1　基于 Gower's 系数的政府记录聚类

政府数据种类较多,同一数据表中可能同时含有数值型属性及分类型属性。常见的相似度计算方法都是针对单一属性的,对于数值型属性,通常使用常见的距离计算方法进行直接度量;对于分类型属性,一般借助分类树进行度量,即需要根据属性值建立准确语义关系。对于政府数据来说,很难通过语义关系建立政府数据分类树,常用的度量方式并不完全满足政府数据的需求。因此,针对政府数据的特点,采用 Gower's 广义相异系数来解决政府混合型数据度量问题。

Gower's 广义相异系数,简称 Gower's 系数,被设计描述不同类型变量下样本间的距离[22]。利用 Gower's 广义相异系数可获得记录间距离,具体实现步骤为:首先,计算各类型变量的均值与标准差,其中分类型变量需根据属性值划分为二分变量(即 0/1 变量)后再进行计算;然后,标准化各类型变量值,数值型属性是对应值减去平均值再除以标准差,分类型属性以同样方式标准化后再乘以补偿系数(补偿 0/1 编码),其中补偿系数为 $1\div\sqrt{2}=0.707\ 1$;最后,根据欧式距离计算标准化后 2 条记录间的距离,其中,欧式距离的计算

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{2}$$

式中,x_i 与 y_i 分别为 2 条记录对应标准化值;n 为属性数量。

利用 Gower's 系数度量记录间距离,结合中心点聚类算法思想就可对政府数据记录进行聚类操作。此过程需要预先设定聚类数目 m,结合 K 匿名对簇中的记录条数的要求,将含有 n 条记录的政府数据聚类数定为 $m=[n/k]$。由于已介绍过中心点聚类算法的实现过程,此处就不再进行赘述。

3.2.2　簇记录数目调整

计算各簇记录数目,当簇中记录数大于 k 时,将距离簇中心较远的记录取出,存于补给数组中备用;当簇中记录数小于 k 时,加入补给数组中距离簇中心最近的记录;若各簇记录数都等于 k 条后,补给数组中仍存在记录,则需分别计算各簇中心点与补给数组记录之间的距离,并将记录分别分派至最近的簇中。

3.3　K 匿名簇共享算法

政府数据在经过基于中心点的记录相似度聚类算法处理后,得到了不少于 K 条记录的簇。然而这些簇并不全满足于 K 匿名的要求,这是因为 K 匿名要求簇中的每条记录是无法区分的。因此,仍需对聚类得到的各簇进行泛化处理。

常见的泛化方式都主要依托于泛化格的构建,同构建分类树一样,政府数据也难以准确构建泛化格,且基于泛化格的泛化方式很容易造成大量的信息损失。比如当前聚类簇在工作类型的属性值为{为私人机构工作,为私人家庭工作},根据泛化格可能将其划分为非政府工作类型,这样无疑加大了信息的损失,使得数据难以再次利用。

因此,需对现有泛化方式进行重新的定义,使其更加适合于政府数据。具体泛化方式为:对于数值型属性,将其最大区间作为泛化值进行共享,即当前簇中该属性的最大值与最小值区间,如[1,25];对于离散型变量,将其属性值集合作为泛化值进行共享,即当前聚类簇中该属性所含值的集合,如{为私人机构工作,为私人家庭工作,为政府部门工作}。

K匿名簇共享算法,不仅需要保证数据符合K匿名的要求,还需要实现政府数据共享功能。因此,K匿名簇共享算法的实现过程主要分为以下2个步骤。

Step 1:簇记录属性泛化。作为基于中心点聚类的K匿名数据共享方法的最后一步,通过前两步已得到了多张初步满足K匿名条件的数据表。K匿名簇共享算法只需按照上述泛化方式对表进行处理,就可得到多张满足K匿名条件的数据表。

Step 2:政府数据共享方式。当政府相关部门根据数据资源目录提出资源共享请求并通过审核后,本部门需要根据其所需属性列,找到对应存储这些属性的K匿名数据表(可能涉及一至多张),并将所查数据覆盖的所有数据表统一共享给所需部门。

4 实验比较分析

4.1 实验数据集

为验证所提算法的有效性,使用Kaggle提供的菲律宾家庭收入支出数据集,对提出的基于中心点聚类的K匿名数据共享方法进行分析。使用Incognito算法作为对比实验,Incognito算法是经典的K匿名隐私保护算法,且其应用范围相当广泛,具有较高的参考价值。

假设以下属性为政府需求部门所需属性列,其中Household Head Sex(户主性别)、Main Source of Income(收入来源)、Household Head Age(户主年龄)、Household Head Class of Worker(工作类型)、Household Head Marital Status(婚姻状态)、Type of Household(房屋类型)都为非敏感属性,Household Head Occupation(户主工作)为敏感属性。上述属性中只有Household Head Age为数值型属性,其余都为分类型属性。

4.2 信息损失量度量

通过计算信息损失量来分析共享数据质量,信息损失程度高则数据质量相对较低。一般情况下,信息损失程度可以通过数据处理前后发生的变化分析。基于中心点聚类的K匿名数据共享算法最终是否可用,在很大程度上取决于该算法造成的信息损失大小。

所提算法涉及的数据属性预处理部分仅仅是对数据表进行划分处理,并未影响属性的使用,而且根据KMedoid-based KADS算法所提出的数据共享方式分析,也不存在属性间的信息损失。因此,即使政府需求部门所需属性列可能分散在共享数据库中的多个表中,仍可以将所需属性列看作一张数据表,统一计算信息损失量。将本文涉及的信息损失进行定义。

设T^*是政府需求部门所需数据表T经过隐私处理得到的匿名数据表,t是匿名表中的一个簇,$|t|$是该簇中含有的记录数,$|T^*|$是匿名数据表含有簇的个数,A代表数值型属性列,B代表分类型属性列。

$a_1,\cdots,a_n$代表簇中含有数值型属性列,$a_1,\cdots,a_n$在簇t上的信息损失为

$$Loss(A)=\frac{\sum_{i=1}^{n}\frac{a_i\max-a_i\min}{|A_i|}}{|t|} \tag{3}$$

式中,$a_i\max$以及$a_i\min$分别为当前数值型属性在簇t上的最大值与最小值;$|A_i|$为当前数值型属性在T^*上的区间差值。

$b_1,\cdots,b_m$代表簇中含有分类型属性列,$b_1,\cdots,b_m$在簇t上的信息损失为

$$Loss(B)=\frac{\sum_{j=1}^{m}\frac{b_j\mathrm{sum}-1}{|B_j|}}{|t|} \tag{4}$$

式中,$b_j\mathrm{sum}$为当前分类型属性在簇t上含有的不同值个数;$|B_j|$为当前分类型属性在匿名数据集上的含

有的不同值个数。因此,对于簇 t 来说,其匿名处理产生的信息损失为

$$Loss(t) = Loss(A) + Loss(B) \tag{5}$$

对于整个匿名数据表 T^* 来说,其信息损失量为

$$Loss(T^*) = \frac{\sum_{i=1}^{|T^*|} Loss(t_i)}{|T^*|} \tag{6}$$

4.3　实验结果分析

图 4 显示了不同 k 值下,2 种算法信息损失情况。根据对比实验可以看出,提出的 KMedoid-based KADS 算法在所有的 k 值下都显示出较小的损失量。这是因为 KMedoid-based KADS 算法保证了每个簇中记录的相似性,降低了泛化程度,提高了数据的可用性。从图中还可以观察出 2 个算法的信息损失程度都随着 k 值的增加而增大,这是因为 k 值数目的提高会导致簇中记录条数增加,从而需要更多泛化处理。

图 5 显示了 $k=15$ 下随非敏感属性数量增加,信息损失量的变化情况。从图中可以观察出,在各个数值下 KMedoid-based KADS 算法都显示出较高的优势。这是因为 Incognito 算法使用的全域泛化方式过于拘谨,造成了较大的信息损失。从图中还可以看出,随着非敏感属性数量的增加,信息损失量的数量也在逐渐升高。这是因为无论是哪一种算法都不可避免地使用泛化技术,非敏感属性数量增多,需要泛化的位置增加,相应的数据可用性也会降低。

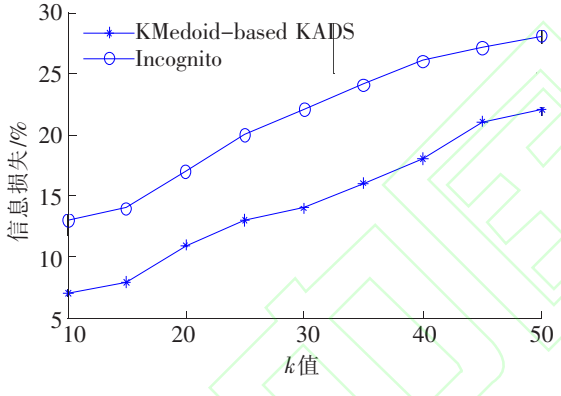

图 4　不同 k 值下的信息损失量

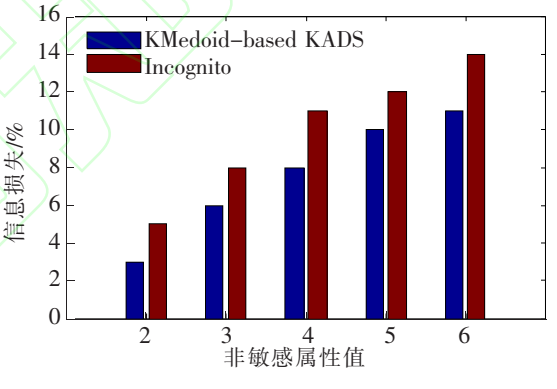

图 5　不同非敏感属性下的信息损失量

5　结论与展望

随着公共数据开放共享的日益普及,公民隐私保护已成为政府数据共享领域关注的热点问题。然而,目前政府在数据共享方面上都存在着不同程度的隐私泄露风险。研究如何在保护公民隐私的基础上有效提高数据的可用性尤为重要。本文构建了基于政务云 + 雾计算模式的隐私保护政府数据共享框架,研究了基于中心点聚类的 K 匿名数据共享方法。通过与经典匿名 Incognito 算法进行实验比较,所提出的 KMedoid-based KADS 算法能有效减少信息损失量,提高数据共享质量。

下一步的研究可以结合政府同部门跨层级资源共享需求,研究在雾层实现跨层级数据共享中的隐私保护总体解决方案,进一步完善基于中心点聚类的 K 匿名数据共享方法,使其可以适应不同的共享需求。

参　考　文　献

[1]Recupero D R, Castronovo M, Consoli S, et al. An innovative, open, interoperable citizen engagement cloud platform for smart government and users' interaction[J]. Journal of the Knowledge Economy, 2016, 7(2): 388-412.

[2]Susanto H, Almunawar M N. Security and privacy issues in cloud-based e-government[M]. Hershey:IGI Global, 2016.

[3]黄如花,刘龙. 英国政府数据开放的政策法规保障及对我国的启示[J]. 图书与情报,2017(1):1-9.

[4]黄如花,刘龙. 我国政府数据开放中的个人隐私保护问题与对策[J]. 图书馆,2017(10):1-5.

[5]国务院办公厅.国务院办公厅关于印发政务信息系统整合共享实施方案的通知[EB/OL].(2017-5-3).[2018-1-31]http://www.gov.cn/zhengce/content/2017-05/18/content_51 94971.htm.
[6]张晓娟,王文强,唐长乐.中美政府数据开放和个人隐私保护的政策法规研究[J].情报理论与实践,2016,39(1):38-43.
[7]刘凌,罗戎.大数据视角下政府数据开放与个人隐私保护研究[J].情报科学,2017,35(2):112-118.
[8]王芳,储君,张琪敏,等.跨部门政府数据共享:问题、原因与对策[J].图书与情报,2017(5):54-62.
[9]黄如花,苗淼.中国政府开放数据的安全保护对策[J].电子政务,2017(5):28-36.
[10]周水庚,李丰,陶宇飞,等.面向数据库应用的隐私保护研究综述[J].计算机学报,2009,32(5):847-861.
[11]董鑫.云计算中数据安全及隐私保护关键技术研究[D].上海:上海交通大学,2015.
[12]邱硕.面向隐私保护的密文数据检索与集合操作的关键技术研究[D].北京:北京交通大学,2017.
[13]何明,常盟盟,吴小飞.一种基于差分隐私保护的协同过滤推荐方法[J].计算机研究与发展,2017,54(7):1439-1451.
[14]Chen R, Mohammed N, Fung B C M, et al. Publishing set-valued data via differential privacy[J]. Proceedings of the VLDB Endowment, 2011, 4(11): 1087-1098.
[15]华佳烽,李凤华,郭云川,等.信息交换过程中的隐私保护技术研究[J].网络与信息安全学报,2016,2(3):28-38.
[16]葛丽娜,张静,刘金辉,等.基于k-同构和局部随机化的隐私保护方法[J].广西师范大学学报:自然科学版,2016,34(4):1-8.
[17]兰丽辉,鞠时光,金华,等.数据发布中的隐私保护研究综述[J].计算机应用研究,2010,27(8):2822-2827.
[18]Chen R, Fung B C M, Mohammed N, et al. Privacy-preserving trajectory data publishing by local suppression[J]. Information Sciences, 2013, 231: 83-97.
[19]Liu F, Li T. A clustering-anonymity privacy-preserving method for wearable iot devices[J]. Security and Communication Networks, 2018,2018:1-8.
[20]Cramér H. Mathematical methods of statistics[M]. Princeton:Princeton University Press, 2016.
[21]Rousseeuw P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational and Applied Mathematics, 1987, 20: 53-65.
[22]Greenacre M, Primicerio R. Measures of distance between samples: noneuclidean[M]// Multivariate Analysis of Ecological Data. [S. l.][s. n.] ,2013: 47-59.

Research on Privacy Protection in Governmental Data Sharing

Shi Yajuan[1], **Piao Chunhui**[1], **Yan Jiaqi**[2]

(1. School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang 050043, China;
2. School of Information Management, Nanjing University, Nanjing 210023, China)

Abstract: With the arrival of big data era, government pays more and more attention to mining the value of government data by means of technology. At the same time, the demand of government for data sharing is increasing. However, there is different degree of privacy leakage risk in the government's data sharing. From the view of privacy protection requirements for data sharing among government departments, this paper proposes an improved K-anonymity data sharing method based on KMedoid clustering (KMedoid-based KADS), according to the characteristics of government data types and complex attributes. Firstly, in order to reduce the sharing range of non-demand attribute data, the original data table is preprocessed and divided into multiple data tables based on attribute correlation. Secondly, the similarity clustering algorithm based on KMedoid clustering is used to deal with the partition. The results show that the data table initially meets the anonymity requirement. Finally, according to the sharing requirements of the data resource request department, an anonymity cluster sharing algorithm is used to provide an anonymous data table for the demand data. Compared with the Incognito algorithm, it is proved that the KMedoid-based KADS algorithm can effectively reduce information loss and improve the usability of shared data.

Key words: government data sharing; privacy protection; fog computing; K-anonymity; cluster