面向大数据的隐私保护方法研究

■ 贺　丹

（东莞理工学院城市学院，广东 东莞 523419）

摘　要：信息技术和网络技术的发展使得大数据成为研究热点，大数据的 4V 特征为人类社会的进步带来了前所未有的机遇和挑战。大数据背景下的隐私泄露问题时有发生，隐私保护问题的解决迫在眉睫。本文针对隐私泄露问题，主要研究面向大数据的隐私保护方法，首先提出了大数据背景下的隐私保护框架，接着归纳总结了隐私保护的关键技术，并以位置隐私保护为例，说明隐私保护技术的具体应用。

关键词：大数据；隐私保护；语法隐私；语义隐私

中图分类号：TP309　　**文献标志码：**A　　**文章编号：**（粤 O）L0150003（2019）01-58-06

一、大数据概述

随着通信技术和互联网技术的飞速发展，新的数据每分每秒都在产生，数据种类和数据规模呈现指数增长的趋势，人们已经进入了大数据时代。被誉为“数据仓库之父”的 Bill Inmon 早在 20 世纪 90 年代就曾提出过“大数据”的概念。《Nature》杂志于 2008 年 9 月推出了“Big Data”专栏，该专栏从互联网、社会经济学、生物医学等方面讨论了大数据时代所面临的挑战 [1]；在工业界，全球最著名的管理咨询公司麦肯锡公司（McKinsey & Company）最早看到了网络平台上海量用户信息潜在的商业价值，并投入大量的人力和财力进行研究，于 2011 年 6 月发布了题为《海量数据，创新、竞争和提高生成率的下一个新领域》的研究报告 [2]，该研究报告对大数据的影响、关键技术、应用领域等方面进行了详尽的阐述，象征着大数据时代的真正来临。

（一）大数据特征

尽管目前学术界和工业界对大数据还没有一个公认的定义，但是随着大数据技术的发展，大数据呈现出一些基本特征，这些基本特征可以总结为 4V 模型 [3]，即数据量大（Volume）、数据种类繁多（Variety）、数据价值密度低（Value）、数据处理速度快（Velocity）。可将大数据的 4V 特征模型表示为图 1。

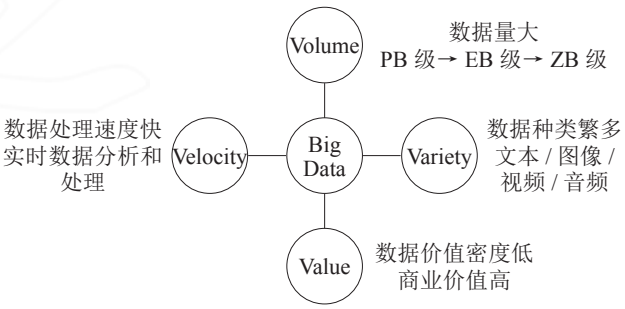

图 1　大数据的 4V 特征模型

从图 1 可知，数据量大是指随着网络技术的发展，数据的数量级已经远远超过了传统的 GB（1GB＝1024MB）或者 TB（1TB＝1024GB），正如国际数据咨询公司（IDC）在报告中指出，预计到 2020 年全球将拥有 35ZB 的数据。数据种类繁多是指随着智能设备种类的逐步增多，数据种类逐渐变得更加复杂，除了最常见的文本数据类型外，还存在图像、视频、音频等数据类型。数据价值密度低是指

收稿日期：2018-08-30

基金项目：东莞理工学院城市学院 2018 年大学生创新创业训练项目“大数据背景下的隐私保护问题研究——以共享单车为例”（项目编号：201813844048）。

作者简介：贺　丹（1992—　），女，湖南湘潭人，东莞理工学院城市学院助教、硕士，主要从事数据质量与数据处理方面的研究。

在数据量呈指数增长趋势的同时，数据中所蕴含的有用信息并没有与数据量的增长呈现相同的趋势，如何从大量数据中挖掘出我们需要的、对研究有价值的信息才是首要解决的问题。数据处理速度快是指随着信息时代的到来，用户对数据的时效性要求逐步提高，需要对数据进行实时分析和处理。

（二）大数据的隐私风险分析

在大数据技术蓬勃发展并得到广泛应用的同时，也面临着一些前所未有的挑战。众所周知，数据在产生的过程中，往往会借助一些载体，这些载体可以是个人或者团体组织。因此，这些数据包含了个人或者团体组织的信息，甚或个人的敏感信息或者隐私信息，数据用户借助这些信息能轻易地识别出个人的身份信息。虽然数据本身没有隐私，但是个人或者团体是有隐私的。

目前，互联网、物联网、移动终端等技术得到了普及，我们的日常生活时时刻刻都在留下数据的痕迹。例如，当我们使用淘宝、京东等购物平台进行消费时，这些购物平台会记录下我们购买物品的信息；当我们使用微信、QQ、微博等聊天软件时，这些聊天软件会记录下我们的交友情况和聊天信息；当我们在使用车载 GPS、百度地图等导航软件时，这些导航软件会记录下我们的位置信息和移动轨迹信息。

当这些第三方平台或者软件公司获取了用户的基本信息之后，他们会利用这些信息挖掘出有价值的信息，从而促进自身业务的发展。这时，用户的基本信息就面临着被泄露的风险，给用户的隐私带来了极大的威胁。一旦发生了数据隐私泄露，将给个人或团体带来经济或者名声上的损失。因此，在我们享受大数据时代所带来便捷的同时，也面临着隐私被泄露的风险。

二、大数据背景下隐私保护框架

为了解决大数据环境下隐私泄露的问题，基于大数据的 4V 模型，结合数据处理的基本流程，本文提出了一种大数据背景下的隐私保护框架。在该框架中涉及了不同的参与者，因此在介绍该框架之前，需要对大数据隐私保护系统中涉及的参与者及其基本操作进行介绍。

正如前文所说，数据的载体是个人或者团体组织，个人或者团体是有隐私的，因此结合大数据的产生、收集、存储、处理、共享、应用等环节，得到如图 2 所示的大数据隐私保护系统中的参与者及其操作。

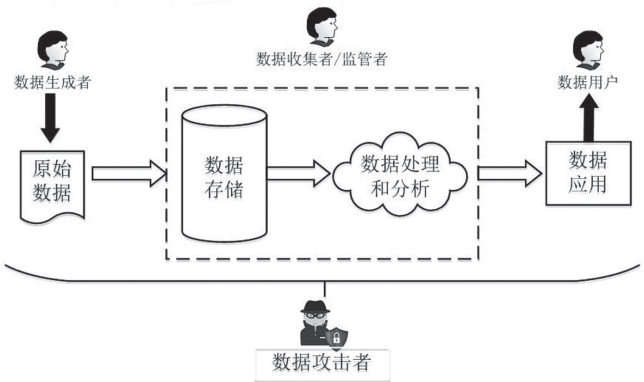

图 2 大数据隐私保护系统中的参与者及其操作

从图 2 可知，数据生成者是指产生数据的个人或者企业，这些生成的数据被传送给数据收集者或数据监管者。例如在使用社交平台时，用户会上传自己的照片，并填写姓名、年龄、联系电话等信息，这些信息往往被社交平台获取，这就为后续的隐私泄露提供了基础。数据收集者或数据监管者负责对数据进行存储、处理和分析，在这些过程中，数据收集者或数据监管者可以从已有数据中挖掘出潜在的价值，他们也可将用户数据买卖给第三方，从而谋取私利，这样数据完全脱离了个人或者企业的管理，发生隐私泄露的风险大大增加。数据用户直接使用数据收集者或数据监管者处理过的数据，数据用户可以查询数据收集者或数据监管者手中的数据。数据攻击者是指为了获取数据生成者的敏感信息的个人或者团体，数据攻击者在数据生成、收集、存储、处理、分析、应用等每一个环节都可能对数据进行攻击，数据攻击者的存在直接造成了敏感信息的泄露。

正是由于数据攻击者的存在，让个人数据随时面临被泄露的风险。因此，结合大数据隐私保护系统中的参与者及其操作，提出一种大数据背景下的隐私保护框架，该框架如图 3 所示。

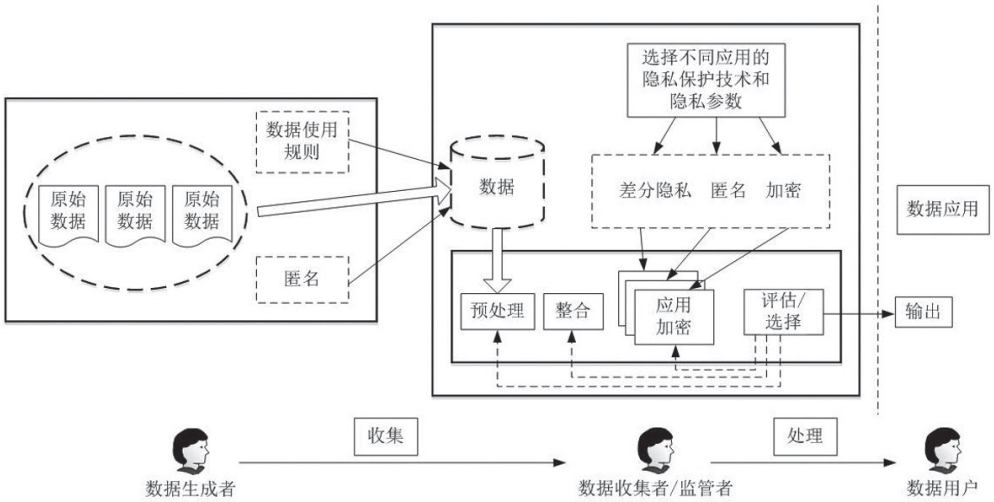

图 3 大数据背景下的隐私保护框架

从图 3 可知，大数据背景下的隐私保护框架主要包括三个部分：数据收集阶段、数据应用阶段、隐私泄露发生阶段。本框架从数据生成者和数据收集者这两个角色的角度进行设计，这两个角色在不同阶段的隐私保护过程中发挥着重要的作用。

在数据收集阶段，数据生成者将自身的数据上传至数据收集者，这些数据为原始数据，可以是购物记录信息、运动轨迹信息、聊天记录信息等。在数据生成者上传数据时，应该指明数据的内容以及数据的用途，此外还可以指明数据的使用权限。在该阶段，数据隐私保护技术主要包括数据使用规则和匿名。例如，在用户使用“滴滴打车”软件时，可以选择使用滴滴平台随机生成的手机号码来代替真实的手机号码，以通过这种方式向滴滴车主隐藏用户的真实手机号码，从而保护用户的个人隐私信息。

在数据应用阶段，数据收集者需要选择不同应用的隐私保护技术和隐私参数对收集到的数据进行隐私保护。常用的隐私保护技术有差分隐私技术、数据匿名化处理、数据加密技术等，不同的隐私保护技术适用于不同的数据类型，也会产生不同的隐私保护效果，在实际的应用场景中，需要根据实际需要来选择合适的隐私保护技术。值得一提的是，数据隐私保护程度越高，数据的可用性就越低，因为随着数据在进行隐私保护之后，数据使用的便利性就降低，从而导致数据的可用性降低。因此对数据进行隐私保护时，需要考虑隐私保护程度和数据可用性之间的关系。

在隐私泄露发生阶段，数据有可能被第三方平台泄露或者被黑客攻击，也就是说隐私泄露可能发生在数据从产生到应用的各个环节。为了降低隐私泄露的风险，可以将数据进行加密处理，即使第三方平台或者黑客获取了数据，也无法得知数据的真实含义，因为经过加密处理之后的数据和原始数据是不相同的，无法直接分析出其真实含义。此外，随着互联网技术的发展，购买网络保险也成为降低数据隐私泄露风险最为直接的手段。

三、大数据背景下隐私保护关键技术研究

大数据背景下隐私保护的关键技术包括：语法隐私保护技术和语义隐私保护技术[4]。其中语法隐私保护技术主要的方法有：k−Anonymity（k− 匿名化）[5]、l−Diversity（l− 多样化）[6]、t−Closeness（t− 保密）[7]，语义隐私保护技术主要的方法有：ε−differential privacy（差分隐私）[8]。

（一）k−Anonymity（k− 匿名化）

k− 匿名化方法是在 1998 年由 Samarati 和 Sweeney 提出的数据加密方法，这种加密方法的主要目的是保证已经公开的数据中包含的个人信息至少有 k−1 条不能通过其他公开信息推断出来，简单地说，就是公开数据中的任意信息相同的组合都至少出现了 k 次。为了更好地说明 k− 匿名化方法的加密过程，以表 1 中某购物网站的用户购物信息为例进行说明。

在该表中，属性可以分成三种类型，第一类是标识属性，一般是作为每条记录的唯一标识，类似于 ID、姓名、电话号码等，这些属性值在数据公开是需要删除的；第二类是半标识属性，类似于性别、年龄、邮编等，这些属性取值是不唯一的，但是往往能够帮助研究人员关联相关数据的标识；第三类是敏感属性，类似于购买偏好、工资等，这些数据往往是有利于研究人员对用户进行行为分析的，也是被直接公开的。

表 1 用户购物信息

姓名	性别	年龄	邮编	购买偏好
李　明	男	24	200083	电子产品
张　超	男	23	200084	洗护用品
刘　红	女	26	200102	护肤品
张　丽	女	27	200104	厨房用具
吴　强	男	36	202208	电子产品
钱明博	男	36	202201	电子产品
胡芳芳	女	34	202218	图书文具
张　慧	女	33	202219	洗护用品

为了说明 *k*− 匿名化加密方法的隐私保护策略，现以 2−Anonymity 为例，对表 1 中的数据进行匿名化保护，得到的数据为表 2 所示。当攻击者获取了表 2 中的数据，想确认用户“李明”的敏感信息“购买偏好”时，通过查询姓名、性别、年龄、邮编这些信息之后，攻击者会发现在表 2 中，至少有两条记录存在相同的性别、年龄、邮编信息，从而无法区分到底哪一条记录才是“李明”的购物信息，这样就保证了用户的敏感信息不被攻击者获取，进而确保用户的隐私不被泄露。

表 2 2−Anonymity 保护用户购物信息

姓名	性别	年龄	邮编	购买偏好
*	男	(20,30]	20008*	电子产品
*	男	(20,30]	20008*	洗护用品
*	女	(20,30]	20010*	护肤品
*	女	(20,30]	20010*	厨房用具
*	男	(30,40]	20220*	电子产品
*	男	(30,40]	20220*	电子产品
*	女	(30,40]	20221*	图书文具
*	女	(30,40]	20221*	洗护用品

从表 2 中可知，*k*− 匿名化方法通过两种方式进行隐私保护。第一种是通过删除属性信息，并用星号“*”代替，如表 2 中删除了姓名属性值，用“*”代替，并将邮编属性的最后一位用“*”代替。第二种方式是将数值型属性值扩展为取值范围，如年龄属性用区间值表示。

（二）*l*−Diversity（*l*− 多样化）

尽管通过 *k*− 匿名化方法可以给用户信息带来一定的安全性，但是由于在进行 *k*− 匿名化处理之后的数据中，同一类记录中的敏感信息缺乏多样性，同样给攻击者留下了攻击漏洞。例如在表 2 中，第 5 行记录和第 6 行记录中性别、年龄、邮编信息都相同，如果攻击者想要获取其“购买偏好”，那么直接可得知其“购买偏好”为电子产品，因为这两条记录在进行了 *k*− 匿名化处理之后，敏感信息属性“购买偏好”值相同，即缺乏多样性。这样就导致经过 *k*− 匿名化处理之后的数据仍无法避免隐私泄露。

为解决上述问题，提出了“多样化”的思想，即在同一类数据中保证敏感信息的取值具有多样性，*l*− 多样化就是“多样化”思想的使用，其中的 *l* 就是指在同一类数据中敏感信息至少有 *l* 种不同的取值。

表 3 中 7 条数据都是同一类数据，因为除了“购买偏好”这一列属性取值不同之外，经过 *k*− 匿名化处理的数据在性别、年龄、邮编这三个属性上的取值都是相同的。从表 3 可看出，前 5 条记录的“购买偏好”属性值都为“电子产品”，第 6 条和第 7 条记录分别为“护肤品”和“家用电器”，即“购买偏好”属性列一共有 3 种取值：电子产品、护肤品、家用电器。这时 *l*− 多样性中的 *l* 取值为 3，这样保证了表中的数据有 3−Diversity 的特点，攻击者获取了性别、年龄、邮编这三个属性值之后，也无法推断出用户的“购买偏好”信息。

表 3 3−Diversity 保护用户购物信息

姓名	性别	年龄	邮编	购买偏好
*	男	(20,30]	20008*	电子产品
*	男	(20,30]	20008*	电子产品
*	男	(20,30]	20008*	电子产品
*	男	(20,30]	20008*	电子产品
*	男	(20,30]	20008*	电子产品
*	男	(20,30]	20008*	护肤品
*	男	(20,30]	20008*	家用电器

通过表 2 和表 3 的分析可知，*k*− 匿名化方法和 *l*− 多样化方法能在一定程度上保护用户的敏感信息不被攻击，进而保护用户的隐私。

（三）*t*−Closeness（*t*− 保密）

通过 *k*− 匿名化方法和 *l*− 多样化方法处理之后的用户信息，仍存在被攻击的可能性，也就是仍存在隐私泄露的风险。为了说明方便，将经过 *k*− 匿名化和 *l*− 多样化处理的数据在表 4 中列出。

表 4 2−Anonymity 和 2−Diversity 处理之后的用户购物信息

姓名	性别	年龄	邮编	工资	购买偏好
*	男	(20,30]	20008*	10k	电子产品
*	男	(20,30]	20008*	10k	图书文具
*	男	(20,30]	10110*	11k	护肤品
*	男	(20,30]	10110*	11k	厨房用品
*	男	(30,40]	10240*	4k	电子产品
*	男	(30,40]	10240*	3k	家用电器
*	女	(20,30]	10210*	9k	护肤品
*	女	(20,30]	10210*	9k	家用电器

表 4 中的敏感属性为“工资”和“购买偏好”，由于表 4 中的 2−Diversity 在处理时并未考虑敏感属性的语义，通过表 4 中第 5 条记录和第 6 条记录可以推断出该用户的工资相对较低，为 4k 或 3k，同时，该用户喜欢购买电子电器产品。

由于 $l-$ 多样化方法在对数据信息进行处理时未考虑敏感属性的语义，才导致用户敏感信息存在泄漏的风险。为了解决表 4 中的问题，提出了 t−Closeness 的概念。t−Closeness 保证了同一类型的记录中，敏感信息分布情况与整体数据敏感信息的分布情况接近（close），并且不超过阈值 t，这里的阈值 t 需要根据实际使用情况进行设置，在本文就不列举。

（四）ε−differential privacy（$\varepsilon-$ 差分隐私）

$k-$ 匿名化、$l-$ 多样化、$t-$ 保密这三种隐私保护的方法都是从语法角度进行数据隐私保护，并且是在数据发布之前对原始数据进行修改。而仅从语法角度进行隐私保护往往不能完全保证用户的信息不被泄露，这就促进了从语义角度进行数据隐私保护方法的发展。与语法角度进行隐私保护的方法不同的是，语义隐私保护通常是在数据查询结果中加入噪声，进行匿名或者模糊处理，从而保护用户数据的隐私。

差分隐私由 Dword 于 2006 年最先提出，是一种充分利用数学理论的、新颖的、稳健的隐私保护技术 [8]。差分隐私技术的提出主要是为了解决下述问题：某购物网站随机发布了 100 名用户的购物偏好数据，其中 10 人偏好购买图书，90 人偏好购买电子产品。攻击者知道了其中 99 人是偏爱购买图书还是偏爱购买电子产品，那么就能利用这 99 人的购物偏好结合购物网站发布的 100 人的购物偏好信息推断出第 100 个人的购物偏好。在这种场景下，攻击者虽未直接获取第 100 个人的购物偏好，但是能从已有信息进行差分攻击。而差分隐私就是为了解决差分攻击而提出的。

差分隐私的思想就是通过一种方法使得查询 99 条记录和查询 100 条记录得到的查询结果相同，这就使得攻击者无法通过对比（即差分）查询结果的差异而推测出第 100 个人的信息。目前，最常用的差分隐私保护方法就是在查询结果中加入随机性，差分隐私方法中的隐私参数 ε 能够严量化隐私保护的程度，从而保证在某一个数据集中插入或者删除一条记录并不会对查询结果造成影响。当攻击者面对通过 $\varepsilon-$ 差分隐私处理之后的数据时，尽管知道除某一条记录之外的其他所有记录，也无法推断出这条记录的敏感信息，这就保证了攻击者无法获取攻击目标的敏感信息，从而达到隐私保护的目的。

由于差分隐私技术在数学理论上有严格的定义标准，尽管数据处理难度较大，但其达到的隐私保护效果是最佳的，因此在实际应用中使用广泛。

四、隐私保护技术应用举例

隐私保护技术是随着互联网通信技术与大数据技术逐步发展起来的，随着人们隐私保护意识的逐步提升，隐私保护技术的应用也得到推广，本文以位置隐私保护为例，说明隐私保护技术的实际应用。

随着移动互联网技术的发展，移动终端的用户群体也在不断扩大，基于位置的服务（Location−based services）应用也更加广泛。例如使用美团外卖点餐时，需要允许 App 访问用户的位置信息，从而推荐附近的餐厅；在使用滴滴打车时，需要允许 App 访问用户的位置信息，从而调度用户附近的车辆；在使用高德地图时，需要允许 App 访问用户的位置信息，从而推荐到达目的地的最快路径。在用户享受“基于位置的服务”为生活带来的便利时，也在承担着位置隐私信息泄露的风险，因为用户的位置信息在提供给服务提供者时，有可能被攻击者截获。

为了减小用户的位置隐私信息被泄露的风险，结合大数据背景下隐私保护的关键技术，研究人员在文献 [9] 中提出了一种用于位置隐私保护的机制，如图 4 所示。

从图 4 可知，在用户使用 App 期间，为了能获得更好的用户体验，会持续获取位置信息，这些位置信息发送给服务提供者，以获取与该位置相关的服务信息。在用户持续使用基于位置的服务时，随着时间和空间的变化，这些变化信息形成了一条与时间和空间相关的有序的时序轨迹，这条轨迹是用户的真实运动轨迹 T。为了防止攻击者从用户的真实运动轨迹 T 推断出用户的真实身份，通过隐私

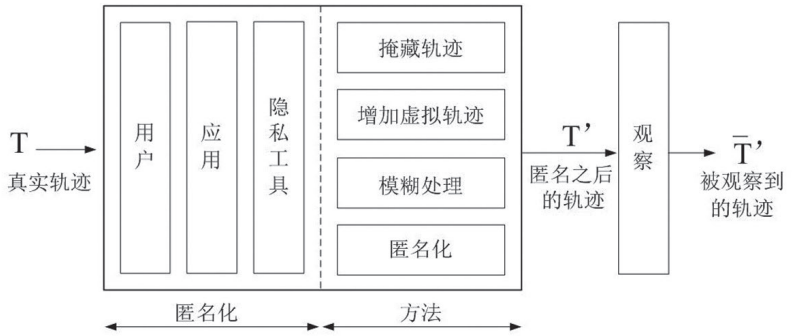

图 4 位置隐私保护机制

保护算法将真实轨迹 T 进行匿名化处理，得到隐私保护技术处理之后的轨迹 T'。攻击者即使获取了匿名之后的轨迹信息 T'，也只能通过观察推测运动轨迹为 T'，而无法推测出用户的真实轨迹 T。

在位置隐私保护机制中的关键技术为位置隐私保护方法，最常用的位置隐私保护方法包括隐藏轨迹、增加虚拟轨迹、模糊处理、匿名化处理等。隐藏轨迹的方法是隐藏用户的轨迹信息，将用户运动轨迹中的部分信息剔除，该方法通常应用在分布式系统中，让用户使用的移动设备在某一段时间段内保持休眠，从而不传递用户的位置信息，这就让用户的位置信息得到隐藏。增加虚拟轨迹的方法是通过人为增加虚拟轨迹来误导攻击者，让攻击者无法区分真实轨迹和虚拟轨迹，这种方法需要保证增加的虚拟轨迹看起来符合常规的用户轨迹，否则无法避免攻击者的攻击。模糊处理的方法是将真实轨迹 T 的时间信息和位置信息进行修改，模糊处理通常都是通过噪声或扰动算法来实现，通过模糊处理之后，轨迹信息中的时间信息和位置信息均不再准确，这给攻击者带来了干扰。匿名化处理是将真实轨迹 T 的标识属性进行修改，从而切断真实轨迹与用户身份之间的关联，修改真实轨迹的标识属性常用的方法就是使用假名来替代真实的用户身份。

五、总结

本文以大数据背景下的隐私保护问题为切入点，首先介绍大数据的 4V 特征模型，并分析大数据背景下的隐私风险。接着为解决大数据背景下的隐私保护问题，提出大数据背景下的隐私保护框架，并对大数据背景下隐私保护的关键技术进行研究，将语法隐私保护和语义隐私保护的方法进行归纳总结。最后以位置隐私保护为例，说明了隐私保护技术的具体应用。

参考文献：

[1] Nature. Big Data [EB/OL]. [2017-02-23] .http://www.nature.com/news/specials/bigdata/index.html.

[2] James Manyika. Big data: The Next Frontier for Innovation, Competition, and Productivity [M] .Chicago: McKinsey Global Institute, 2011.

[3] 田铁刚 . 大数据的特点及未来发展趋势研究 [J]. 无线互联科技 ,2018,15 (09) :61-62.

[4] Sabrina D C D V, Foresti S, Livraga G, et al. DATA PRIVACY: DEFINITIONS AND TECHNIQUES [J] .International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2012,20 (06) :793-817.

[5] Thooyamani K P. Protecting Privacy When Disclosing Information: K Anonymity and Its Enforcement through Suppression [J]. Ijcoa Com, 1970:1-19.

[6] Machanavajjhala A, Kifer D, Gehrke J. L-diversity: Privacy beyond k-anonymity [J] .Acm Transactions on Knowledge Discovery from Data, 2007,1 (1) :3.

[7] Li N, Li T, Venkatasubramanian S. T-Closeness: Privacy Beyond k-Anonymity and l-Diversity [C] //IEEE, International Conference on Data Engineering. IEEE, 2007:106-115.

[8] Dwork C, Mcsherry F, Nissim K. Calibrating noise to sensitivity in private data analysis [C] //Conference on Theory of Cryptography. Springer-Verlag, 2006:265-284.

[9] Shokri R, Freudiger J, Hubaux J. A Unified Framework for Location Privacy [J] .Epfl, 2010.

[责任编辑：蹇　柯]