第5章 人工智能数据隐私保护方法

5.1 数据脱敏

数据脱敏(Data Masking)是指从原始环境向目标环境进行敏感数据交换时,通过一定的方法消除原始环境中数据的敏感性,并保留目标环境业务所需的数据特性或内容的数据处理过程[1]。常见的敏感数据包含个人姓名、联系电话、家庭住址、银行账户详情、身份证明编号等。数据脱敏技术对这类敏感信息实施了有效的处理策略,确保其在流通过程中安全,同时在不违背规范使用原则的前提下,尽可能地保护了数据的价值和隐私。同时,数据隐私作为人民的合法权益,我们应当以服务人民的爱国精神对其进行保护。

在数据流通过程中,数据脱敏会创建真实的数据副本,以保留原始数据的真实性、完整性和统计特性,同时保护数据免受恶意行为或公开披露的影响。数据脱敏一般经过3个步骤,分别是脱敏数据识别、数据脱敏方案制定和执行、效果对比。首先,读入原始数据,然后根据原始数据特点,扫描并识别出疑似敏感数据。在此基础上,用户根据实际需求对疑似敏感数据制定相应的脱敏方案并执行。脱敏后,用户需检查前后两个版本的数据,判断是否符合预期。此过程需要确保脱敏的成本可控,并生成符合业务需求的数据结果。除此之外,数据脱敏还是一种数据访问控制形式,涉及数据替换操作,使得未授权用户不能查看实际值,从而安全可靠地处理了数据。

5.1.1 数据脱敏类型

1. 静态数据脱敏

静态数据脱敏(Static Data Masking, SDM)是在敏感数据存储或共享之前对其应用一组固定脱敏规则的过程。该方法通常用于不经常变更或在一段时间内保持不变的数据。如图5-1所示,在医疗领域,医疗记录中包含了患者的个人信息、诊断结果、治疗方案等敏感数据,这些信息需要在医疗研究、统计分析和共享数据时得到保护。因为医疗记录通常是静态的,一旦记录,就不会发生过大改变,这就需要对这些数据进行静态数据脱敏处理。但是,静态数据脱敏需要复制原

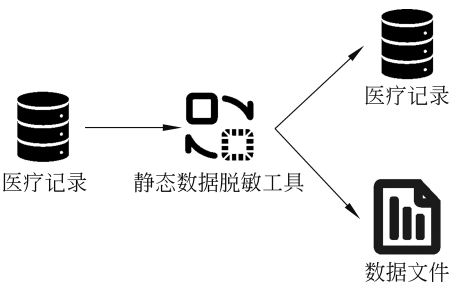

图5-1 医疗记录静态数据脱敏

始数据库，这可能非常耗时且成本高昂。

2. 动态数据脱敏

动态数据脱敏(Dynamic Data Masking，DDM)是对敏感数据实时进行脱敏。当不同用户访问或查询敏感数据时，它会动态调整脱敏规则更改数据。动态数据脱敏主要应用于在客户支持或病例处理等应用程序中实现基于角色的数据安全。如图 5-2 所示，在医疗案例中，护士需要查看患者的病历信息以提供护理服务，但并不需要直接访问患者的个人身份信息；医生进行诊断和制定治疗方案时需要访问患者更多的医疗数据，如历史病例。动态数据脱敏可以保证不同角色的用户获知不同的数据，同时保护敏感数据不被泄露。

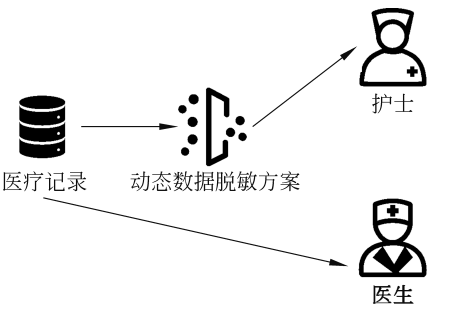

图 5-2　医疗机构动态数据脱敏

5.1.2　数据脱敏技术

数据脱敏的目的是保护敏感数据的隐私和安全，同时确保数据在需要共享、处理或传输时仍具有一定的可用性和有效性。通过数据脱敏，可以最大限度地减少敏感数据被未经授权地访问或滥用的风险，从而保护个人隐私和遵守相关的法律法规。通常，数据脱敏有以下 5 种技术。

(1) **数据仿真**：根据原始敏感数据的特定规则，生成与原始数据具有相同属性和关联关系的新数据。这些新数据经过脱敏处理后，仍然保持原始数据的编码和校验规则，确保了数据的可用性和有效性。举例来说，经仿真处理后的姓名仍然是有意义的姓名，如张三变换为李四，地址数据也仍然是有效的地址信息。这种方法能在保护数据隐私的同时，保持数据的实用性和业务关联性。

(2) **数据替换**：将敏感数据的原始内容替换为相似但虚构的伪造敏感数据，破坏了数据的可读性。替换后的数据将不再保留原有的语义和格式，而是被转换成特殊字符、随机字符或者其他固定值字符。

(3) **数据加密**：通过加密算法，如哈希算法、同态加密等，对敏感数据进行加密，将数据转换为不可读的格式，只有拥有密钥的授权用户，才能访问原始数据。加密技术提供了更高等级的隐私保护，但会影响数据的可用性。

(4) **数据截取**：对原始数据中的部分内容进行截断，只保留其中一部分信息，而丢弃其他信息，但需要确保截断后的数据仍然具有足够的可用性和有效性。

(5) **数据混淆**：无规则打乱敏感数据的内容，以隐藏敏感信息。使用这种方式，需对数据的敏感字段进行重新排列或进行其他方式的打乱，使得原始数据的结构得以保留，但其中的具体敏感信息却无法被轻易识别。

5.1.3　数据脱敏的应用场景

目前，政府、企业和机构的数据平台中存储着大量敏感数据。根据现行法律法规，为了确保数据安全，在数据的采集、传输、交换和共享过程中必须采取必要的措施来防止数据泄

露。因此，数据脱敏技术应用于数据共享的场景中，并且在许多领域都有广泛的应用。

1. 政务行业

公安、税务等政务部门采集了大量公民的个人信息，部门之间的数据共享存在不同级别的访问权限，例如，公安部门的大量个人信息不能直截了当地传递给其他部门，或是对公众开放。因此，政府行业需要对所拥有的信息进行数据脱敏处理，以对敏感部分进行不可逆处理，并降低数据在共享过程中被重新聚合分析的风险。

2. 医疗行业

医院存储了患者的隐私信息，不法分子可能通过收买医院工作人员或第三方维护和开发人员窃取患者的隐私数据以获得高昂的潜在利用价值。为了满足国家对医疗数据隐私保护的要求，同时有效保护用户的隐私数据，维护和提升医疗卫生领域的形象和公信力，部署数据脱敏产品对患者的敏感数据进行脱敏是非常必要的。

3. 金融行业

金融领域如银行，涉及个人信息、账号密码、交易记录等，一旦被恶意人员窃取，将会对整个金融系统造成不可挽回的后果。因此，金融机构需要制定严格的数据脱敏策略，在进行风险评估、市场分析等业务活动时尽可能保护客户的隐私。

习题

习题 1：数据脱敏有哪些类型？有什么不同？

参考答案：

静态数据脱敏和动态数据脱敏。静态数据脱敏是离线进行脱敏，需要复制原始数据库，耗时；动态数据脱敏是实时完成动态脱敏，根据用户角色进行不同程度的脱敏，逻辑框架更为复杂。

习题 2：数据脱敏有哪些技术？

参考答案：

仿真、替换、加密、数据截取、数据混淆。

习题 3：请联系生活，举一些生活中数据脱敏的例子。

参考答案：

例如，选举后公示期间将展示个人信息，其中姓名栏中间或最后一位的名字被替换；身份证中间位置通常不展示，等等。

5.2 数据匿名化

我们在做什么，在哪里工作，在哪里生活，看什么医生……我们的个人数据在不断被收集、存储和使用。随着收集和存储的数据量不断增加，在未经他人知情或同意的情况下，个人信息被访问和滥用的风险逐渐增大，对我们的生活造成严重困扰，甚至是威胁。例如，由于个人相关信息泄露，我们总是接到莫名其妙的骚扰电话。因此，当一个组织不可避免地使用我们的数据时，我们更希望他们永远无法定位追溯到我们，这就是数据匿名化的本质。从

敏感数据脱敏的角度看，数据匿名化技术和数据脱敏的目标是相同的。但与数据脱敏不同，数据匿名化更注重对个人数据的深度处理，以确保数据无法被还原出原始信息。

数据匿名化(Data Anonymization)是指为了保护与该数据相关的人的隐私，从数据集中隐藏或删除个人可识别的过程[2]。数据匿名化使得其他人无法从数据中链接或识别到特定的个人，同时保持信息用于测试、分析或第三方共享等的有效性。数据匿名化通常会保留尽可能多的数据，并且匿名数据往往与原始数据相似。例如，当收集了完整的出生日期时，可以通过隐藏具体的日期并仅保留年份来实现匿名，从而避免个人身份信息暴露。

公布匿名化的数据时，需要防止身份、属性和推断这 3 种类型的信息披露[3]。身份泄露是指通过链接匿名数据中的特定记录，从而揭露个人的身份信息[4]。例如，用 A 代替张三，B 代替李四，C 代替王五，如果 A、B 和 C 的信息被链接到外部数据源中的特定个人，就可能导致身份泄露；属性泄露是有关个人的数据公开时，属性会相应被泄露。例如，部门员工公示的匿名记录中显示所有年龄超过 40 岁的员工已经获得工资、奖金。如果知道某员工是 43 岁并且在这个部门工作，就知道他已经获得奖金这一属性，即使该员工的记录在匿名员工记录中无法与其他人区分；推理泄露是指攻击者将数据与其他数据集关联来获取有关个人的隐私信息。当匿名数据发布时，可能导致通过推理间接披露其他相关敏感数据的情况发生。

5.2.1 数据匿名化基本方法

匿名化技术能在一定程度上减少数据集中的原始信息，一般有泛化、抑制、扭曲、交换 4 种基本方法，且不同方法之间可以组合实现。下面通过一个具体的例子理解这 4 种方法，其原始数据详细信息见表 5-1。

表 5-1 原始数据详细信息

序号	姓名	地　址	邮编	用电量/kW
123	张三	广东省广州市北京路 117 号	35400	434
234	李四	江苏省徐州市南京路 14 号	51400	289
456	王五	安徽省滁州市杭州路 23 号	81200	45
789	赵六	四川省广元市上海路 89 号	68100	872

(1) **泛化**：泛化是指用更广义的内容替换特定的内容，用于向未授权方隐藏个人信息，如数值、地址等。例如，在表 5-1 中，可以对用电量和地址信息进行泛化处理，将具体的用电量“434”泛化为“400-500”，将具体的地址“广东省广州市北京路 117 号”泛化为“广东省广州市北京路中段”，等等。

(2) **抑制**：抑制是指将数据集中的部分数据的值更改为没有意义的值，如用“******”替换原始数据[5]。抑制通过删除信息来隐藏数据，原始数据中存在的记录完全从输出中删除。例如，在表 5-1 中，我们可以对个人姓名进行抑制处理，将“张三”替换为“**”，以隐藏真实信息，等等。

(3) **扭曲**：扭曲是指一种将数据转变为其他形式的过程，只有被授权的人才能查看原始数据。例如，在表 5-1 中，通过对称加密算法等技术将“地址”转换为无法读取的数据，只

有拥有密钥的用户才能解密并查看该数据。

(4) **交换**：交换是指将真实的数据值替换为虚构但相似的数据值。此方法可应用于整个数据集，也可应用于数据库中的特定字段或属性。例如，在表 5-1 中，可以对姓名进行随机生成，将“张三”交换为“Alice”，等等。

5.2.2　数据匿名化技术

在匿名数据发布过程中，为了防止个人身份和敏感信息被识别，我们需要关注匿名化后的结果。下面介绍 3 种匿名化效果的量化评估方式，以便数据发布者能更好地衡量匿名化处理的效果，从而更好地保护个人隐私。

1. K-匿名

K-匿名(K-anonymity)需要确保没有一个人的信息可以与同一数据集中至少 $K-1$ 个其他人区分。换言之，对于任何给定的记录，数据集中至少有 K 个记录，其中所有属性特征的值都相同[6]，这增加了从数据集中直接筛选出记录并攻击的难度。这意味着，攻击者不能仅通过查看识别属性的值识别数据集中的某个特定的人，因为数据集中至少还有其他人具有完全相同的值。例如，在表 5-2 中，攻击者通过邮编和年龄可以定位一条记录，但经过 K-匿名后，对邮编和年龄做抑制，即使攻击者掌握了具体的邮编和年龄，也无法确定用户患上何种疾病。

表 5-2　原始患者医疗记录

序　　号	邮　　编	年　　龄	疾　　病
1	47677	29	心脏病
2	47602	22	心脏病
3	47678	27	心脏病
4	47905	43	流感
5	47909	52	心脏病
6	47906	47	癌症
7	47605	30	心脏病
8	47673	36	癌症
9	47607	32	癌症

K-匿名技术的工作原理是相似的分组组合，并泛化或抑制包含识别信息的数据字段。此方法防止个人信息被披露，也防止个人在数据集中被识别，使团体组织更容易保护消费者、员工等个人群体的隐私，尤其是与第三方共享或公开数据。

但在使用 K-匿名保护敏感信息时，用户要意识到它的局限性和潜在的弱点。首先，它不能防止外部因素、附加信息、数据链接来重新识别个人的攻击，无法防止属性信息的公开，导致其无法抵抗同质攻击、背景知识攻击、补充数据攻击等。相关攻击介绍如下。

(1) **同质攻击(Homogeneity Attack)**：某包含 K 个条目的数据组别中，每个条目的敏感属性值完全相同。在这种情况下，攻击者能轻易识别并提取对应的敏感信息。如表 5-3 所

示，邮编为 476 开头的，年龄为 20～30 岁的患有心脏类疾病。

表 5-3　*K*-匿名化医疗记录示例

序　　号	邮　　编	年　　龄	疾　　病
1	476**	2*	心脏病
2	476**	2*	心脏病
3	476**	2*	心脏病
4	4790*	≥40	流感
5	4790*	≥40	心脏病
6	4790*	≥40	癌症
7	476**	3*	心脏病
8	476**	3*	癌症
9	476**	3*	癌症

（2）**背景知识攻击（Background Knowledge Attack）**：即使在某个包含 K 个条目的数据组别中，每个条目对应的敏感属性值不完全相同，持有先验知识的攻击者也能利用其所掌握的信息，高概率地推断并揭露隐私数据。如表 5-3 所示，如果攻击者知道某个人的邮编开头是 476，年龄为 30 岁及 30 岁以上，且在日本，而日本地区的心脏疾病发病率很低，就可知道这个人患有癌症。

（3）**补充数据攻击（Complementary Release Attack）**：匿名公开数据存在多种类型，且匿名方法不同，那么攻击者可以通过关联这些公开数据推测用户信息。

此外，随着 K 值的增大，数据隐私保护更好，同时也会导致数据的效用降低，如何抉择适当的 K 值将变得非常困难。因此，使用 K-匿名评判时，需要根据具体的任务和案例实现。

2. *L*-多样性

L-多样性（L-diversity）是指在基于敏感属性的情况下，任何记录的信息无法与数据集中的至少 L 个其他记录区分[7]，弥补了 K-匿名的固有弱点。在 K-匿名的基础上，L-多样性要求敏感数据必须具备多样性，进而保证用户的隐私不能通过同质、背景知识等攻击推测出来。如表 5-4 所示，任何邮编开头为 476，年龄为 20～30 岁的人群患上不同疾病，有“心脏病”“骨折”“流感”，在这种情况下，攻击者也无法通过同质、背景知识等攻击推测出来。

表 5-4　*L*-多样性匿名化医疗记录示例

序　　号	邮　　编	年　　龄	疾　　病
1	476**	2*	心脏病
2	476**	2*	骨折
3	476**	2*	流感
4	4790*	＞＝40	流感

续表

序　　号	邮　　编	年　　龄	疾　　病
5	4790*	＞＝40	心脏病
6	4790*	＞＝40	癌症
7	476**	3*	心脏病
8	476**	3*	癌症
9	476**	3*	癌症

与 K-匿名一样，L-多样性也不能保证完全的隐私保护。例如，攻击者可以将异常高的发生率的属性联系起来，表 5-4 中，邮编开头为 476，年龄为 30 岁及 30 岁以上，有 2/3 的人患有癌症，那么这类人群大概率患有心脏病，此类攻击称为偏斜攻击（Skewness Attack）。再者，L-多样性并没有考虑敏感值的语义，相似的语义可以令攻击者学习到更多重要的信息，这类攻击称为相似攻击（Similarity Attack）。此外，为了满足 L-多样性，数据集会删除许多记录，导致信息丢失，因为它不仅必须识别和保护敏感属性，且只有当数据集中每个属性至少有 L 个不同的值时，才能实现 L-多样性。

3. T-相近性

T-相近性（T-closeness）是对 L-多样性的进一步细化处理，在对属性处理时还增加了对属性数据值的考虑，进而防止匿名数据遭受相似攻击。T-相近性要求数据集中敏感属性的统计分布与整个数据的敏感信息分布接近，不超过阈值 T[8]。例如，在表 5-5 中，薪水这一敏感属性与整个数据的分布是接近的，因为在表格中所有人的薪水都是不同的，同时每条记录代表不同的人，在这种情况下，攻击者就无法通过相似攻击的方式窥探到特定人群的隐私。

表 5-5　T-相近性匿名化医疗记录示例

序　　号	邮　　编	年　　龄	薪　　水	疾　　病
1	4767*	＜＝40	3k	胃溃疡
3	4767*	＜＝40	5k	胃癌
8	4767*	＜＝40	9k	肺炎
4	4790*	＞＝40	6k	胃炎
5	4790*	＞＝40	11k	流感
6	4790*	＞＝40	8k	支气管炎
2	4760*	＜＝40	4k	胃炎
7	4760*	＜＝40	7k	支气管炎
9	4760*	＜＝40	10k	胃癌

T-相近性是一个更严格的隐私保护要求，它不仅要求对敏感属性进行识别和保护，还要求匿名化后的数据集中的敏感属性分布与总体数据集中的敏感属性分布相似。相比之下，T-相近性相对于 K-匿名或 L-多样性来说更难以实现，因为它要求匿名化后的数据不仅

要保持匿名化效果，还要保持数据分布的相似性。然而，实现 T-相近性也增强了数据的隐私保护效果，使得匿名化后的数据更符合实际应用需求，同时保护了个人隐私。

5.2.3 数据匿名化技术的应用场景

随着人工智能和机器学习技术的不断涌现，需要大量数据来训练模型，然后在不同的业务领域之间共享。数据匿名化提供了与数据共享相关的隐私问题的方案，使其实际上不可能从数据集中重新识别个人信息。下面介绍数据匿名化的实际应用场景。

1. 媒体通信

电信公司和媒体公司使用数据匿名化保护敏感信息，如通话/消息日志、位置详细信息和个人信息。二者共享并使用匿名数据进行报告、研究和分析，而不用担心会损害客户隐私。同时，数据匿名化可以让电信公司发现网络流量的模式和瓶颈，而不会涉及个人隐私。这样，公司可以优化网络资源分配，改善用户体验，提高网络性能。

2. 医疗卫生

医疗保健行业使用数据匿名化保护患者的隐私，同时允许他们的数据用于合法目的，如分析、报告和研究。在医疗保健领域，数据匿名化可以保护病史、个人信息和治疗细节。例如，匿名化后的数据可用于评估某些药物治疗效果的研究，或用于确定疾病暴发的趋势，而不暴露患者信息。

3. 金融服务

金融服务公司，如银行、券商和保险公司，采用数据匿名化保护敏感信息，如财务历史、个人信息和交易信息。例如，匿名化后的数据可用于识别欺诈模式，或测试营销活动的有效性，而不暴露任何可识别信息。

习题

习题 1：数据匿名化有哪些基本类型？

参考答案：

泛化、抑制、扭曲、交换。

习题 2：数据匿名化有哪些检验技术？

参考答案：

K-匿名、L-多样性、T-相近性。

习题 3：K-匿名是什么？有哪些缺陷？

参考答案：

K-匿名是指对于任何给定的记录，数据集中至少有 K 个其他记录，其中所有属性特征的值都相同。缺陷：①不能防止外部因素、附加信息、数据链接来重新识别个人的攻击，无法防止属性信息的公开，导致其无法抵抗同质攻击、背景知识攻击、补充数据攻击，等等；②随着 K 值的增大，会导致数据的效用降低，如何抉择适当的 K 值将变得非常困难。

习题 4：L-多样性是什么？有哪些缺陷？

参考答案：

L-多样性是指在基于敏感属性的情况下，没有一条记录的信息可与数据集中的至少 L

个其他记录区分。缺陷：①L-多样性会遭受偏斜攻击、相似攻击；②为了满足 L-多样性，数据集会删除许多记录，导致信息丢失。

5.3 差分隐私

5.3.1 差分隐私的形式化定义

1. 传统隐私保护方法的不足

差分隐私的提出源于对传统隐私保护技术的反思。传统的隐私保护方法，如数据脱敏、匿名化[9]等，虽然可以一定程度上保护个人隐私，但在大数据时代，这些方法面临严重的局限性。大量研究表明，通过辅助信息或背景知识的攻击，攻击者仍然可以获取到敏感数据。例如，若仅对数据进行去标识化处理，攻击者可以利用关联攻击，推断或识别出发布数据中的个体信息。在 1997 年美国马萨诸塞州保险委员会(Massachusetts Committee of Insurance)医疗数据泄露事件中，委员会删除了发布数据中的所有用户身份标识，包括姓名、地址和社会安全号码(SSN)，只保留出生日期、性别和邮政编码等个人信息，然而麻省理工学院(MIT)的研究生 Latanya Sweeney 通过将发布数据与该州的选民数据集进行比对，如图 5-3 所示将医疗记录与选民数据集相匹配，最终找到了时任州长 William Weld 的医疗记录，这意味着去标识化的方式存在着巨大的漏洞。

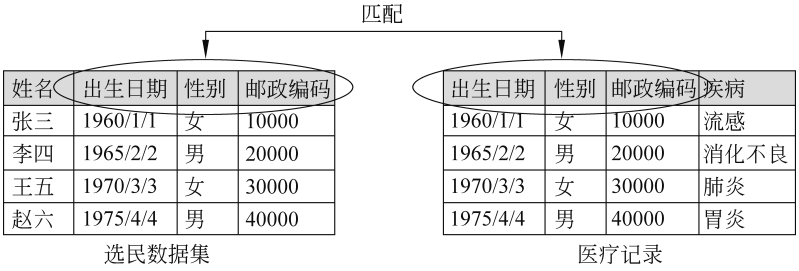

图 5-3　将医疗记录与选民数据集相匹配

Latanya Sweeney 后续提出了 K-匿名技术，确保可以将用户个人敏感信息隐藏在人数至少为 K 的群体中。然而，K-匿名技术也存在不足，无法抵御同质攻击、背景知识攻击、补充数据攻击等，并且当数据集存在异常值时，K-匿名的最优泛化十分困难。另外，其对隐私的保护程度并没有严格定义。

2. 差分隐私方法的提出

为了克服传统隐私保护方法中的缺陷，Dwork 等[10]提出了差分隐私的概念。差分隐私通过对数据或模型添加噪声，防止隐私信息泄露。具体来说，差分隐私要求在数据集中加入或删除一个个体的数据时，数据发布的结果应该在一定范围内保持不变。如图 5-4 所示，相邻数据集 D 和 D' 仅相差一条数据(数据集 D' 比数据集 D 少了一条用户 F 的数据)，攻击者分别对两个数据集提出查询男性总人数的请求，如果直接给出查询的结果，则 D 回答为 3，D' 回答为 2，如果攻击者已知 D 和 D' 仅相差一条用户 F 的数据这一背景知识，他就可以推断出“F 是男性”这一隐私信息；如果使用差分隐私机制对查询结果添加噪声进行扰动，则 D

和D'的输出分布几乎是不可区分的，这样攻击者就无法推断用户 F 的隐私信息。

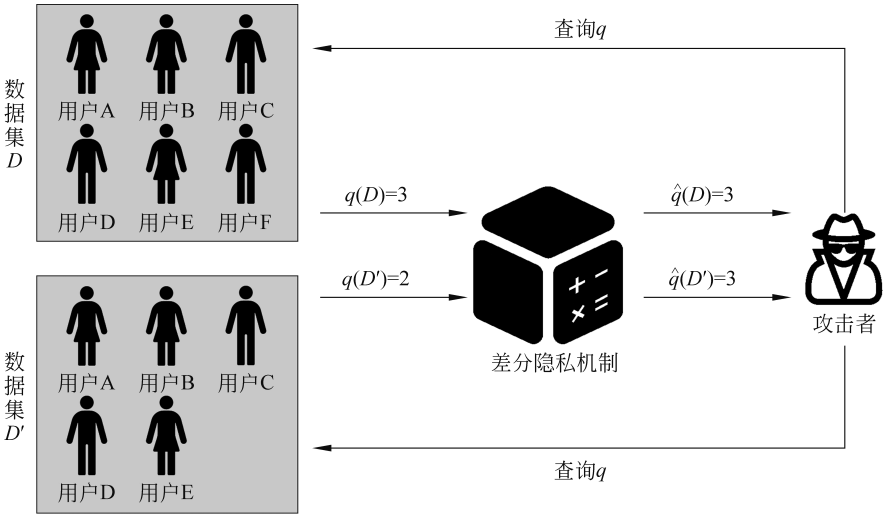

图 5-4　差分隐私示例

与数据脱敏、匿名化等传统隐私保护技术相比，差分隐私的优势在于可以量化分析隐私的保护程度，不受攻击者先验知识的影响，可以在理论上提供强大的隐私保证，并且只需要非常轻量级的运算与存储开销。

3. 差分隐私的定义[11]

1）ε-差分隐私

ε-差分隐私的数学定义[11]如下。

定义 5-1（ε-差分隐私）：对于定义域为$\mathcal{D}$、值域为 R 的随机机制 $M:\mathcal{D}\to R$，若对于任意两个相邻输入数据集 D，$D'\in\mathcal{D}$以及任意输出子集 $S\subseteq R$，都满足：

$$\Pr(M(D)\in S)\leqslant e^{\varepsilon}\cdot\Pr(M(D')\in S) \tag{5-1}$$

则称随机机制 M 的输出满足ε-差分隐私。

① 随机机制 M：一个随机化算法，特点是对于一个特定的输入，输出服从某一分布（如高斯分布）的随机值。

② 相邻数据集（Neighboring Dataset）：两个汉明距离为 1（即只相差一条数据）的数据集。

③ ε：隐私预算（Privacy Budget），描述了单行数据对于整个数据集所包含信息的影响的上界，也是采用差分隐私机制时所允许的隐私泄露程度的上界。隐私预算的大小直接影响差分隐私保护程度和数据实用性。较小的隐私预算可以提供更强的隐私保护，但同时会引入更多的噪声，降低数据的效用；相反，较大的隐私预算虽然可能降低隐私保护的强度，但噪声的干扰会减少，从而提高数据的效用。因此，在确定隐私预算时，必须根据实际情况，在保护隐私和保持数据效用之间找到平衡点。

差分隐私的核心思想如下：当数据集 D 和 D'为相邻数据集且数据 x 仅在数据集 D 中时，即使攻击者知道数据集内除 x 外所有元素的信息，也依然无法确定当前数据集为 D 还是为 D'，也就无法确定 x 是否存在于当前数据集中，更无法获得 x 的具体内容，从而保护