Анонімізація даних. Використання відкритого програмного продукту Amnesia

Чуканова Світлана, канд. пед. н., Виконувачка обов'язків директора Наукової бібліотеки НаУКМА

У сучасному світі дані стали одним з найцінніших ресурсів для наукових досліджень.

Однак робота з даними, особливо персональними, вимагає ретельного дотримання етичних принципів та правових норм.

Ефективна анонімізація даних

Що таке анонімізація?

Анонімізація - це процес обробки персональних даних таким чином, щоб особу неможливо було ідентифікувати. Це не просто видалення імен або очевидних ідентифікаторів, а комплексний процес зниження ризику ідентифікації.

Мета ефективної анонімізації:

Зниження ризику ідентифікації до можливого мінімуму

Правова безпека: забезпечити, щоб оброблені дані не підпадали під дію законодавства про захист персональних даних (деперсоналізація, анонімізація)

Збереження корисності даних для дослідницьких цілей

Захист прав та свобод осіб, дані яких обробляються

Основні виклики анонімізації:

Баланс між захистом та корисністю: чим більше ми захищаємо дані, тим менше вони можуть бути корисними для досліджень

Технологічний розвиток: нові технології можуть зробити раніше анонімні дані ідентифікованими

Контекстуальність: те, що є анонімним в одному контексті, може не бути таким в іншому

Концепція ідентифікованості

Ідентифікованість - це можливість визначити, що інформація стосується конкретної особи. Згідно з законодавством про захист даних, персональні дані визначаються як:

> "будь-яка інформація, що стосується ідентифікованої або ідентифікованої живої особи"

Розширене розуміння ідентифікованості:

- **Не лише ім'я**: особа може бути ідентифікованою навіть без знання її імені
- **Контекстуальність**: те, що ідентифікує в одному контексті, може не ідентифікувати в іншому
- **Непрямі ідентифікатори**: комбінація, здавалося б, безневинних характеристик

Прямі ідентифікатори:

Ім'я та прізвище

Номер паспорта

Номер телефону

Email адреса

Непрямі ідентифікатори:

Дата народження

Поштовий індекс

Професія

Рідкісні захворювання

Комбінація демографічних характеристик

Практичний приклад:

Уявіть датасет з інформацією про пацієнтів лікарні, де видалено імена, але залишено:

- Вік: 45 років
- Стать: чоловік
- Район проживання: Печерський
- Діагноз: рідкісне генетичне захворювання
- Дата госпіталізації: 15 березня 2024

Навіть без імені, така комбінація характеристик може дозволити ідентифікувати пацієнта.

Спектр ідентифікованості

Замість простого бінарного підходу "ідентифіковано/анонімно", ідентифікованість існує у вигляді спектру:

```
Прямо ідентифіковано ⟷ Непрямо ідентифіковано ⟷ Малоймовірно ідентифіковано ⟷ Неможливо ідентифікувати
```

```
[Персональні дані]
```

```
[Анонімні дані]
```

Фактори, що впливають на положення на спектрі:

Специфіка обробки: чутливість змінних, техніки анонімізації

Середовище даних: технічні та організаційні заходи контролю доступу

Управління ризиками: як ідентифікуються та пом'якшуються ризики

Динамічність спектру:

Технологічний розвиток: нові методи аналізу даних можуть зсунути дані в бік більшої ідентифікованості

Доступність додаткових даних: нові публічні датасети можуть дозволити зв'язування

Зміна контексту: використання даних для іншої мети може змінити рівень ризику

Три ключові індикатори ідентифікованості

Для оцінки ефективності анонімізації використовуються три основні критерії:

Виділення (Singling Out)

Визначення: Можливість ізолювати записи про конкретну особу в датасеті.

Як це працює:

Якщо ви можете відокремити деякі або всі записи про особу, то вона є "виділеною"

Навіть якщо ви не збираєтеся нічого робити з цією інформацією, сам факт можливості виділення робить особу ідентифікованою

Практичний приклад:
У датасеті співробітників університету є запис:

- Посада: "Професор квантової фізики"
- Вік: 67
- Стаж: 40 років

Якщо в університеті є лише один такий професор, то цей запис легко "виділити".

Методи оцінки ризику виділення:

Аналіз "багатства" даних (кількість атрибутів)

Оцінка рідкісності комбінацій характеристик

Врахування розміру референтної популяції

Зв'язуваність (Linkability)

Визначення: Можливість поєднати кілька записів про одну й ту ж особу або групу осіб.

Ефект мозаїки: Окремі джерела даних можуть здаватися неідентифікуючими окремо, але в комбінації дозволяють ідентифікацію.

Типи зв'язування:

Внутрішнє зв'язування: в межах одного датасету

Зовнішнє зв'язування: між різними датасетами

Темпоральне зв'язування: зв'язування записів у часі

Практичний приклад:

Датасет А: Анонімні медичні записи з поштовим індексом

Датасет Б: Публічні записи про власників нерухомості

Датасет В: Соціальні мережі з геолокацією

Комбінація цих трьох джерел може дозволити ідентифікацію пацієнтів.

Методи протидії зв'язуванню:

Маскування: заміна або видалення ключових змінних

Токенізація: заміна ідентифікаторів випадковими токенами

Генералізація: заміна точних значень діапазонами

Придушення: видалення записів з рідкісними комбінаціями

Виведення висновків (Inferences)

Визначення: Можливість вивести, вгадати або передбачити деталі про особу.

Типи виведення:

Статистичне виведення: базується на кореляціях у даних

Машинне навчання: використання алгоритмів для передбачення

Експертне знання: використання спеціалізованих знань

Практичний приклад:
Із датасету покупок у аптеці можна вивести:

Вагітність (покупка вітамінів для вагітних)

Діабет (регулярні покупки тест-смужок)

Депресія (комбінація ліків)

Джерела додаткової інформації для виведення:

Публічні дані (перепис населення)

Соціальні мережі

Знання експертів (лікарів, вчителів)

Сімейні зв'язки

Правові аспекти оцінки ризиків

Принцип "засобів, розумно ймовірних для використання"

Згідно з Recital 26 GDPR, при визначенні ідентифікованості особи слід враховувати:

"всі засоби, розумно ймовірні для використання, такі як виділення, контролером або іншою особою для прямої або непрямої ідентифікації особи"

Об'єктивні фактори для врахування:

Економічні витрати:

Скільки коштує процес ідентифікації?

Чи є ці витрати розумними для потенційного зловмисника?

Часові витрати:

Скільки часу потрібно для ідентифікації?

Чи є цей час прийнятним?

Технічні можливості:

Які технології доступні зараз?

Як вони можуть розвинутися в майбутньому?

Правові обмеження:

Які існують правові бар'єри для доступу до додаткової інформації?

Тест "мотивованого зловмисника"

Концепція тесту

Тест мотивованого зловмисника - це методологія оцінки ризиків ідентифікації, яка встановлює реалістичний рівень загрози.

Профіль мотивованого зловмисника:

Базові характеристики:

Розумно компетентна особа (не експерт, але й не новачок)

Має доступ до загальнодоступних ресурсів

Володіє базовими дослідницькими навичками

Має конкретну мотивацію для ідентифікації

Доступні ресурси:

Інтернет-пошук

Публічні бази даних

Бібліотеки та архіви

Соціальні мережі

Недорогі підписні сервіси

Дослідницькі техніки:

Перехресна перевірка інформації

Аналіз патернів

Соціальна інженерія (без кримінальних дій)

Комбінування різних джерел

Що НЕ входить в профіль:

Глибокі технічні знання (хакерство)

Доступ до спеціального обладнання

Значні фінансові ресурси

Готовність до кримінальних дій

Інсайдерська інформація

Типи мотивацій:

1. **Особиста вигода**:
    - Фінансова вигода
    - Шантаж
    - Промислове шпигунство
2. **Шкода або помста**:
    - Підрив довіри до організації
    - Особисті образи
3. **Журналістські розслідування**:
    - Пошук "новинних" історій
    - Розкриття інформації про публічних осіб
    - Суспільний інтерес
4. **Політичні або активістські цілі**:
    - Кампанії проти організацій
    - Політична боротьба
    - Соціальні рухи
5. **Наукова цікавість**:
    - Демонстрація вразливостей
    - Академічні дослідження
    - Тестування систем
6. **Випадкове впізнання**:
    - Знайомі чи родичі
    - Колеги
    - Сусіди

Практичне застосування тесту:

Крок 1: Аналіз датасету

Які змінні присутні?

Наскільки рідкісні комбінації характеристик?

Які можливі шляхи зв'язування?

Крок 2: Оцінка мотивації

Хто може бути зацікавлений у цих даних?

Наскільки сильна мотивація?

Які можливі шкоди від ідентифікації?

Крок 3: Аналіз доступних ресурсів

Яка додаткова інформація доступна публічно?

Які техніки можуть бути використані?

Наскільки реально отримати необхідну інформацію?

Крок 4: Оцінка ймовірності успіху

Чи зможе мотивований зловмисник ідентифікувати особу?

Скільки часу та ресурсів це займе?

Наскільки впевнено можна ідентифікувати?

Джерела інформації для ідентифікації

Категорії попередніх знань:

1. Професійні знання:

Лікар може впізнати свого пацієнта в анонімному медичному дослідженні

Вчитель може ідентифікувати учня в освітньому датасеті

HR-менеджер може впізнати співробітника в опитуванні

2. Сімейні та особисті зв'язки:

Члени сім'ї знають медичну історію один одного

Друзі обізнані про особисті обставини

Сусіди спостерігають за повсякденним життям

3. Громадські зв'язки:

Колеги по роботі

Однокласники

Члени спільнот

Публічно доступна інформація:

1. Офіційні реєстри:

Виборчі списки

Земельний кадастр

Реєстр підприємств

Судові рішення

2. Онлайн-ресурси:

Соціальні мережі (Facebook, Instagram, LinkedIn)

Професійні профілі

Форуми та блоги

Новинні сайти

3. Архівні матеріали:

Газетні архіви

Університетські щорічники

Професійні публікації

Історичні записи

4. Комерційні бази даних:

Генеалогічні сайти

Адресні довідники

Професійні каталоги

Маркетингові бази

Методи збору інформації:

1. Пасивний збір:

Пошук у Google

Перегляд соціальних мереж

Вивчення публічних записів

2. Активний збір:

Опитування знайомих

Розміщення оголошень

Соціальна інженерія

3. Технічні методи:

Аналіз метаданих

Використання API соціальних мереж

Автоматизований збір даних

Оцінка різних типів доступу до даних

1. Публічний (Open Data)

Характеристики:

- Дані доступні будь-кому
- Неможливо контролювати використання
- Складно відкликати дані
- Максимальний ризик

Вимоги до анонімізації:

- Найвищий рівень захисту
- Врахування всіх можливих сценаріїв атак
- Консервативний підхід до ризиків
- Ретельне тестування

2. Доступ для визначеної групи

Характеристики:

- Обмежена кількість отримувачів
- Можливість контролювати доступ
- Договірні зобов'язання
- Середній рівень ризику

Додаткові заходи:

- Договори про нерозголошення
- Технічні засоби контролю доступу
- Моніторинг використання
- Можливість відкликання

3. Внутрішнє використання

Характеристики:

- Використання в межах організації
- Максимальний контроль
- Мінімальний ризик витоку
- Можливість швидкого реагування

Особливості:

- Менш жорсткі вимоги до анонімізації
- Можливість збереження більшої корисності
- Внутрішні політики безпеки

Динамічна природа ризиків та необхідність перегляду

Фактори, що змінюють ризики:

1. Технологічні зміни:

- Нові алгоритми машинного навчання
- Збільшення обчислювальної потужності
- Розвиток методів де-анонімізації
- Покращення алгоритмів зв'язування

2. Зміни в доступності даних:

- Нові публічні датасети
- Витоки даних з інших джерел
- Зміни в політиці конфіденційності платформ
- Нові джерела інформації

3. Соціальні зміни:

- Зміна ставлення до приватності
- Нові форми активізму
- Політичні зміни
- Зміна мотивацій зловмисників

Стратегії моніторингу:

1. Технологічний моніторинг:

- Відстеження наукових публікацій
- Участь у конференціях
- Моніторинг PoC атак
- Консультації з експертами

2. Моніторинг середовища:

- Відстеження нових датасетів
- Аналіз витоків даних
- Моніторинг соціальних мереж
- Оцінка змін у законодавстві

3. Внутрішній моніторинг:

- Регулярні аудити
- Тестування на проникнення
- Моніторинг доступу
- Аналіз інцидентів

Періодичність перегляду:

Високий ризик (щомісячно або щоквартально):

- Медичні дані
- Фінансові дані
- Дані про дітей
- Дані про вразливі групи

Середній ризик (щорічно):

- Освітні дані
- Дані опитувань
- Демографічні дані

Низький ризик (раз на 2-3 роки):

- Агреговані статистичні дані
- Історичні дані
- Публічні дані