УДК 004.056.53
Т. Сачик, Н. Загородна
(Тернопільський національний технічний університет імені Івана Пулюя)

ЗАХИСТ ПЕРСОНАЛЬНОЇ ІНФОРМАЦІЇ В ЗАДАЧАХ АНАЛІЗУ ТА ОБРОБКИ ВЕЛИКИХ ДАНИХ

UDC 004.056.53
T. Sachyk, N. Zagorodna
(Ternopil Ivan Puluj National Technical University, Ukraine)

PROTECTION OF PERSONAL INFORMATION IN THE OBJECTIVES OF ANALYSIS AND PROCESSING OF BIG DATA

Інформаційні технології не лише полегшують наше повсякденне життя, але й збирають та зберігають величезні обсяги приватної інформації користувачів. І, якщо частина цих даних має цілком загальний характер, то інші персональні дані, включаючи прізвища людей, дати народження, номера страхових полісів і рахунків, дозволяють ідентифікувати особу. Часто компанії можуть публікувати результати своїх досліджень або ж передавати зібрані дані стороннім особам (науково-дослідним центрам) для аналізу. Необхідність публічного поширення або ж передача третій стороні приватних даних поставила нові виклики щодо захисту. Деякі організації, включаючи Інститутську наглядову раду (IRB, США) і Європейське агентство по оцінці лікарських засобів (EMA), вимагають, щоб дослідники і компанії-виробники ліків анонімізували свої дані, перш ніж публікувати результати своїх досліджень, з метою захисту персональних даних їх учасників і їх права на недоторканність приватного життя. Як результат, публікація даних, що зберігають конфіденційність, стала активною дослідницькою сферою. Отже, існує необхідність пошуку інструментів анонімізації персональних даних з метою мінімізації негативних наслідків можливого порушення їх конфіденційності.

Анонімізація, або редагування даних – процес видалення або приховування персональних даних з метою їх подальшого використання. На перший погляд, вирішення цієї проблеми видається доволі тривіальним: адже достатнього просто видалити стовпці, що містять прямі ідентифікатори, такі як імена та номери соціального страхування тощо. Тим не менше, було доведено, що такого підходу недостатньо для збереження конфіденційності. Ця проблема виникає тому, що все ще можливо поєднувати різні набори даних або мати базові знання про людей, щоб зробити висновки про особу. Повторна ідентифікація особи досягається за допомогою зв’язування атрибутів, відомих як квазі-ідентифікатори (QID), таких як стать, дата народження або поштовий індекс. Науковці з США довели, що поєднуючи відкриту інформацію з різних джерел можна однозначно ідентифікувати 70-90% людей.

Існує кілька моделей, які пропонують формальні гарантії щодо захисту конфіденційності особи при публікації даних. Зосередимось на *k*-анонімізації, оскільки на відміну від інших моделей (ℓ-різноманіття, *t*-близькість та диференційна конфіденційність), які мають обмеження в використанні, ця модель є простою для розуміння і базовою у багатьох сферах використання. Більше того автори [1] вказують на актуальність моделі k-анонімності як основи для побудови більш надійних моделей.

У моделі *k*-анонімізації кожна людина представлена у вигляді набору атрибутів, включаючи QID –атрибути, які можуть бути пов'язані із зовнішньою інформацією з метою однозначної ідентифікації особи. Захист конфіденційності в методі *k*-анонімізації полягає в тому, щоб гарантувати, що кожен набір QID відображається принаймні в *k* записах у наборі даних, або, що дані будь-якої конкретної особи не відрізняються від даних принаймні *k* - 1 інших осіб щодо QID. Мета методу полягає в тому, щоб зробити квазі-ідентифікатори неточними та менш інформативними. *k*-анонімізація зазвичай досягається шляхом узагальнення та приховування даних (наприклад, опусканням імен осіб і заміною п'ятизначних поштових індексів лише їх першими двома цифрами) з метою створити класи еквівалентності, що мають однакові QID. Тому метою нашого дослідження є порівняння найбільш відомих методів k-анонімізації (Datafly, Incognito, Mondrian) з огляду на використання ресурсів та корисність залишкових даних.

1. B. Kenig and T. Tassa. A Practical Approximation Algorithm for Optimal k-Anonymity. Data Min. Knowl. Discov., 25(1):134–168, 2012.