ІДЕНТИФІКАТОРИ ПЕРСОНАЛЬНИХ ДАНИХ ТА ПОБУДОВА ПУБЛІЧНОГО НАБОРУ ДАНИХ

Подоляка О.О.[1], Подоляка О.М.[2]
[1]Харківський національний університет імені В. Н. Каразіна, м. Харків
[2]Національний аерокосмічний університет імені М.Є. Жуковського «ХАІ», м. Харків

Розкриття особистостей людей у «знеособлених» даних називається реідентифікацією або деанонімізацією. В індустрії захисту персональних даних виділяють наступні типи ідентифікаторів: прямі або явні; непрямі або квазіідентифікатори; конфіденційні або сенситивні та ключові. Регламент GDPR (General Data Protection Regulation) зобов'язує видавця запобігти всіляким ризикам розкриття персональних даних у відкритих наборах даних.

Початковим етапом побудови безпечного публічного набору є розробка сенситивної моделі чи моделі конфіденційних даних. Обов’язковим етапом побудови публічного набору даних є анонімізація прямих ідентифікаторів. National Institute of Standards and Technology (NIST) визначає анонімізацію, як процес, який видаляє зв’язок між ідентифікуючим набором даних та суб’єктом даних [1]. На наступному етапі захисту даних необхідно зменшити точність значень або грануляцію непрямих ідентифікаторів.

Слід зазначити, що існують моделі атак проти приватності, що не потребують реідентифікації, а саме, моделі журналіста та маркетолога [2]. У роботі [3] розглянуто моделі атак, що ґрунтуються на розкритті значень непрямих ідентифікаторів, а також моделі оцінки ризиків відповідних загроз.

Існують дві основні техніки для зменшення точності оцінок ідентифікаторів або грануляції. Зменшення грануляції можна розглядати як процедуру усунення відмінностей схожих квазіідентифікаторів. Можна сказати, що дана процедура виконує розбиття вихідної таблиці на кластери шляхом об’єднання схожих записів (наприклад, близьких за віком, вагою, поштовим кодом тощо). Ці кластери також називають класами еквівалентності. Кожному класу відповідає множина конфіденційних даних. Ця стратегія називається «сховатися в натовпі». Під натовпом в даному випадку розуміється множина невиразних об’єктів, кожен з яких ховає свої таємниці в цьому натовпі.

Література:

1. Fung B., Wang ke, Wang L., Debbabi M. A framework for privacy-preserving cluster analysis. Conference: Intelligence and Security Informatics, 2008. С. 46 - 51.

2. Marques, Joana & Bernardino, Jorge. Analysis of Data Anonymization Techniques. In Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 2020. С. 235-241. ISBN: 978-989-758-474-9.

3. Podoliaka O., Mushkatblat V., Kaplan A. Privacy Attacks Based on Correlation of Dataset Identifiers: Assessing the Risk, 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), 2022. С. 0808-0815. ISBN: 9781665483032.