DOI: https://doi.org/10.26565/2304-6201-2024-61-07

УДК 004.8

Подоляка Оксана Олександрівна
к. т. н., доцент
Харківський національний університет імені В. Н. Каразіна,
площа Свободи 4, Харків, Україна, 61022
e-mail: podoliaka@karazin.ua;
https://orcid.org/0000-0002-3401-2996

Подоляка Олексій Миколайович
старший викладач
Харківський національний університет імені В. Н. Каразіна,
площа Свободи 4, Харків, Україна, 61022
e-mail: alex.podolyaka@gmail.com;
https://orcid.org/0000-0002-5755-3728

Оцінка корисності публічного набору даних для аналітичних досліджень

Організації та агенції публікують різні дані, які призначені для аналізу, навчання систем штучного інтелекту та інших дослідницьких цілей. Відповідно до прийнятих регуляцій у сфері захисту персональних даних публічні дані мають бути знеособлені та захищені від різних загроз розкриття персональних даних. Усунення цих загроз реалізується шляхом зменшення точності даних під час підготовки публічних даних. Втрата точності, вочевидь, призводить до зменшення корисності даних для аналізу. У роботі розглядаються ентропійні метрики корисності та проблеми їх обчислюваності, а також метрики втрати корисності окремих підмножин публічних даних.

Мета. Розробка ефективних метрик оцінки корисності публічного набору даних для аналізу з урахуванням вимог захисту персональних даних.

Методи дослідження. Інформаційна безпека, теорія інформації Шенона, управління даними (Data Governance).

Результати. Запропоновані метрики оцінки втрат інформації та корисності даних для аналізу на основі ентропійних метрик теорії інформації Шеннона. Запропоновано процедури, спрямовані на підвищення швидкодії обчислень розглянутих метрик.

Висновки. Описано процедури побудови безпечного публічного набору даних. Розглянуто питання застосування ентропійних метрик теорії інформації Шеннона для оцінки втрат інформації та корисності даних для аналізу. Показано, що обчислення зазначених метрик є складною, практично не здійсненною для великих баз даних, обчислювальною задачею. Запропоновано процедури, спрямовані на підвищення швидкодії обчислень розглянутих метрик. А саме, створення менш точної копії вихідних даних та формування випадкової вибірки із великої бази даних для обчислення необхідних статистик. Розглянуто метрики оцінки корисності для окремих підмножин (кластерів) публічних даних.

Ключові слова: *конфіденційність, деідентифікація, публікація даних, корисність даних, GDPR (General Data Protection Regulation).*

Як цитувати: Подоляка О. О, Подоляка О. М. Оцінка корисності публічного набору даних для аналітичних досліджень. *Вісник Харківського національного університету імені В. Н. Каразіна, серія Математичне моделювання. Інформаційні технології. Автоматизовані системи управління*. 2024. вип. 61. С.61-67. https://doi.org/10.26565/2304-6201-2024-61-07

How to quote: Podoliaka O. O., Podoliaka O. M., “Assessing the utility of a public dataset for analytical research”, *Bulletin of V. N. Karazin Kharkiv National University, series Mathematical modelling. Information technology. Automated control systems*, vol. 61, pp.61-67, 2024. [In Ukrainian]. https://doi.org/10.26565/2304-6201-2024-61-07

Вступ

Публічні дані містять чутливу для людей інформацію, яка може бути використана для суспільно значущих цілей, наприклад, аналітичними агентствами для складання прогнозів, керівництвом компаній для організації ефективного управління, фінансовими організаціями для визначення оптимальних бізнес-стратегій, моделювання процесів, аналізу вразливостей тощо. Важливо розуміти, що публічні дані можуть бути використані зловмисниками для злочинних цілей, таких як: шпигунство, шантаж, переслідування, здирство тощо.

Тому персональні дані мають бути деідентифіковані у публічних наборах даних. Для цього, зазвичай, застосовується шифрування, додавання статистичних шумів, узагальнення даних із

© Подоляка О. О, Подоляка О.М., 2024
This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0.

використанням таксонномій, зменшення гранулярності (точності оцінок) тощо. Очевидно, що методи деідентифікації знижують корисність даних для аналізу [1-4]. Детальні огляди методів деідентифікації можна отримати з [5-7]. Слід зазначити, що жоден з цих методів не дає повної безпеки ідентифікації людей у публічному наборі даних.

1 Ідентифікатори персональних даних та побудова публічного набору даних

Розкриття особистостей людей у «знеособлених» даних називається реідентифікацією або деанонімізацією. Реідентифікація розкриває конфіденційні дані у деідинтифікованих наборах публічних даних. В індустрії захисту персональних даних виділяють наступні типи ідентифікаторів.

Прямі ідентифікатори або явні ідентифікатори – атрибути даних, які можуть безпосередньо ідентифікувати особу. Наприклад: ID, паспорт та номер соціального страхування, ім'я – прізвище тощо. Згідно з прийнятими регуляціями, явні ідентифікатори повинні бути видалені з публічного набору даних [8-10].

Непрямі ідентифікатори або квазіідентифікатори – загальнодоступні атрибути даних, які безпосередньо не можуть ідентифікувати людину, але можуть використовуватися для ідентифікації людей у різних моделях атак. Наприклад, вік, поштовий індекс, дата народження тощо.

Конфіденційними або сенситивними ідентифікаторами називатимемо підмножину непрямих ідентифікаторів, що містять важливу особисту чи конфіденційну інформацію. Вони мають велику цінність як для зловмисників, так і для дослідників. Ці дані несуть безпосередню загрозу приватності та, у разі розкриття, можуть завдати відчутної шкоди. Наприклад, дані про здоров’я, зайнятість, освіту, матеріальне становище, релігію, різного роду уподобання тощо.

Ключовими називаються ідентифікатори, які зберігають ключі, що зв’язують таблиці публічного набору даних або посилання на інші дані (документи, фото, відео тощо).

Регламент GDPR (General Data Protection Regulation) зобов'язує видавця запобігти всіляким ризикам розкриття персональних даних у відкритих наборах даних. Основні ризики пов’язані з легкою доступністю для зловмисників обсягів персональних даних, достатніх для ідентифікації людей у знеособлених публічних даних. Хакери також можуть використовувати техніки аналізу даних для визначення кореляцій, що несуть загрозу приватності.

Початковим етапом побудови безпечного публічного набору є розробка сенситивної моделі чи моделі конфіденційних даних. Цю модель формують:

1) підмножини прямих, непрямих та конфіденційних ідентифікаторів набору даних;

2) модель конфіденційності, яка визначає критерії безпеки конфіденційних даних;

3) модель корисності даних, яка встановлює обмеження втрати інформації в ході деідентифікації набору даних.

Обов’язковим етапом побудови публічного набору даних, згідно з усіма прийнятими у світі регуляціями, 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 [11]. Зазвичай, анонімізація полягає у видаленні прямих ідентифікаторів (телефон, ідентифікаційний код, адреса тощо), або заміщенні їх синтетичними даними, що обчислюються згідно деякого алгоритму.

На наступному етапі захисту даних необхідно зменшити точність значень або грануляцію непрямих ідентифікаторів. Це робиться, щоб завадити зловмиснику реідентифікувати людей, коли йому відомі точні значення даних.

Слід зазначити, що існують моделі атак проти приватності, що не потребують реідентифікації, а саме, моделі журналіста та маркетолога [12, 13]. Зокрема, мета журналіста – зіпсувати репутацію видавця даних, а мета маркетолога – адекватне маркетингове дослідження. У роботі [14] 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. Перша – додавання до даних випадкових шумів, а друга – узагальнення значень на основі ієрархій чи таксономій. Наприклад, маскування даних зірками або синтетичними значеннями можна вважати екстремальною формою узагальнення.

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.

2 Оцінка корисності даних або інформаційних втрат

Для оцінки корисності даних у ході деідентифікації пропонується використовувати математичні моделі теорії ймовірностей та теорії інформації Шеннона [15, 16]. Ці моделі тісно взаємопов'язані. Можна сміливо сказати, що фундаментом теорії інформації є теорія ймовірностей.

Для початку розглянемо загальну модель оцінки корисності, на підставі якої буде побудовано прикладну математичну модель оцінки корисності окремих кластерів.

Нехай джерело повідомлень X – це множина значень певного набору даних. Тоді кількість інформації окремого повідомлення $x, x \in X$ обчислюється за формулою Хартлі.

$$I(x) = -\log p(x) \tag{1}$$

де $I(x)$ показує скільки біт інформації несе повідомлення x, ймовірність якого становить $p(x)$.

Ентропія $H(X)$ – це середня кількість інформації, що надходить на одне випадкове повідомлення з джерела повідомлення X.

$$H(X) = -\sum_{x \in X} p(x) \cdot log\big(p(x)\big) \tag{2}$$

Якщо є залежність між повідомленнями або елементами множин $x \in X$, $y \in Y$, то спільна кількість інформації визначається за формулою.

$$I(x, y) = -\log p(x, y) \tag{3}$$

$$p(x, y) = p(x) \cdot p(x/y) \tag{4}$$

де $p(x, y)$ і $p(x/y)$ – спільна ймовірність і умовна ймовірність подій x, y, $p(x, y) \le p(x/y)$, $p(x, y) \le p(x)$.

Нехай X та Y – множини значень ідентифікаторів деякого публічного набору даних. Маємо на увазі, хоча це не обов'язково, що X – квазіідентифікатор, а Y – конфіденційний ідентифікатор (наприклад, X – вік пацієнта, а Y – його захворювання). Встановити наявність залежності між X та Y можна оцінивши обсяг спільної інформації цих множин $I(X, Y)$. Вона обчислюється на основі безумовної $H(X)$ та умовної ентропії $H(X/Y)$.

Припустимо, що є залежність між елементами множин $x \in X$ та $y \in Y$ деякого кластера набору даних. Тоді спільна ентропія множин X, Y визначається за формулою.

$$H(X, Y) = -\sum_{\substack{x \in X \\ y \in Y}} \big[p(x, y) \cdot log\big(p(x, y)\big)\big] \tag{5}$$

$H(X, Y)$ також називають ентропією об'єднання двох джерел, вона показує середню кількість інформації, що припадає на два випадкові повідомлення (x, y) джерел X, Y.

$$H(X, Y) = H(Y, X) = H(X) + H(Y/X) = H(Y) + H(X/Y) \tag{6}$$

$H(X/Y)$ – загальна або повна умовна ентропія об'єднання X та Y, або кількість інформації джерела X без урахування спільної інформації, що міститься в X та Y.

$$H(X/Y) = -\sum_{\substack{x \in X \\ y \in Y}} \big[p(x, y) \cdot log\big(p(x/y)\big)\big] \tag{7}$$

Загальна умовна ентропія використовується для обчислення втрат інформації джерела X з урахуванням залежності джерел X та Y.

Спільна інформація джерел X то Yвизначається за формулою.

$$I(X,Y) = I(Y,X) = H(X) - H(X/Y) = H(Y) - H(Y/X) \quad (8)$$

На рисунку 1 показано взаємозв'язок спільної ентропії, умовної ентропії та спільної інформації.

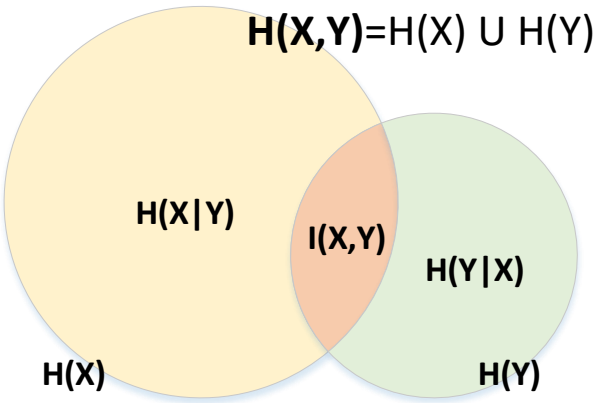

Рис. 1 Спільна, умовна ентропія та спільна інформація

Кореляцію множин X, Y можна визначити, як відношення обсягів інформації $H(X/Y)$ та $H(X)$, $0 < H(X/Y) \leq H(X)$.

$$R(X/Y) = 1 - \frac{H(X/Y)}{H(X)} \quad (9)$$

$$R(Y/X) = 1 - \frac{H(Y/X)}{H(Y)} \quad (10)$$

Область визначення R перебуває у діапазоні $[0; 1]$, тобто R відображає залежність підмножин даних.

Розглянемо переваги і недоліки моделей (6 – 8).

До переваг розглянутих моделей корисності можна віднести простоту інтерпретації оцінок, отриманих згідно з формулами (1-10). Тобто, легко зрозуміти скільки порядків точності даних (бітів) було втрачено після деідентифікації. Однак, ентропійні метрики, обчислені на основі великих обсягів даних, втрачають адекватність через екстремальне усереднення інтегральних результатів та їх обчислення є важким завданням. Ми розглянули лише два домени таблиці (X, Y). Якщо розглядати загальний випадок, коли $X \epsilon X^*; Y \epsilon Y^*$, де X^*– множина всіх квазіідентифікаторів, а Y^*– множина всіх конфіденційних ідентифікаторів у формули (6-8) слід додати суми за цими множинами. Оскільки обчислення звичайних ймовірностей має лінійну складність, а умовних – квадратичну, для таблиць з мільйонами записів і сотнями доменів обчислення розглянутих оцінок перетворюється на нездійсненну обчислювальну задачу.

Для вирішення описаних вище проблем пропонується два підходи.

1. Формування менш точного набору даних шляхом зменшення грануляції або точності метричних оцінок атрибутів початкових даних.

2. Формування випадкової репрезентативної вибірки даних задля підвищення швидкості обчислення необхідних статистик.

Ці підходи вирішують проблеми обчислюваності ентропійних метрик великих баз даних.

Перший підхід. Неточний тимчасовий набір даних необхідний виключно, щоб швидко порахувати необхідні ймовірності для метрик корисності. Важливо розуміти, що для аналізу здебільшого не потрібна екстремальна точність. Наприклад, якщо база містить історії транзакцій за рахунками клієнтів, то навіщо маркетологу знати історії платежів до останнього долара або

цента. Унікальність значень ідентифікаторів – це пряма загроза приватності. Наприклад, за сумою платежу в знайденому чеку хакер може з великою ймовірністю ідентифікувати особистість в базі, з якої видалені прямі ідентифікатори.

Другий підхід. Слід підкреслити, що в теорії інформації обчислення розглянутих ентропій виконується в повній групі подій. Це означає, що необхідні ймовірності мають обчислюватися по всій базі даних. Оскільки отримання необхідних статистик у повній групі подій неможливе, пропонується обчислити необхідні ймовірності на підставі репрезентативної вибірки невеликого розміру. У базі даних записи зазвичай упорядковані за деяким індексом, тому отримання випадкової вибірки з великої бази являє собою складну обчислювальну задачу. Ефективні методи отримання випадкової вибірки розглянуті в роботах [17-19].

Важливо розуміти, що на вході кластеризації значення даних різних рядків таблиці спотворюються не рівномірно. Отже, втрати інформації окремих елементів даних і кластерів публічного набору даних можуть відрізнятися. Тому потрібна метрика корисності для окремого кластера та груп кластерів. Ентропійні метрики кластера, обчислені на основі статистики всієї таблиці неможливо застосувати для кластерів, оскільки ентропії мають обчислюватися для повної групи подій. Тому, для оцінки втрат інформації або корисності окремого кластеру чи підмножин кластерів пропонується використовувати відношення кількості інформації деідентифікованого кластера до кількості інформації вихідної підмножини даних. Дану метрику легко порахувати, її також можна використовувати в якості одного з критеріїв кластеризації у процесі побудови публічного набору даних.

Висновки.

Описано процедури побудови безпечного публічного набору даних. Розглянуто питання застосування ентропійних метрик теорії інформації Шеннона для оцінки втрат інформації та корисності даних для аналізу. Показано, що обчислення зазначених метрик є складною, практично не здійсненною для великих баз даних, обчислювальною задачею. Запропоновано процедури, спрямовані на підвищення швидкодії обчислень розглянутих метрик. А саме, створення менш точної копії вихідних даних та формування випадкової вибірки із великої бази даних для обчислення необхідних статистик. Розглянуто метрики оцінки корисності для окремих підмножин (кластерів) публічних даних.

СПИСОК ЛІТЕРАТУРИ

1. Dankar, F., Emam, K.E., Neisa, A., Roffey, T.: Estimating the re-identification risk of clinical data sets. BMC Med. Inform. Decis. Mak, 2012. №12 (66). С. 1-15.
2. B.A. Malin, D. Karp, R.H. Scheuermann. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. J. Investig. Med., 2010. 58 (1). С. 11-18.
3. Li, Tiancheng & Li, Ninghui. On the tradeoff between privacy and utility in data publishing. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009. С. 517-526.
4. Fung, Benjamin & Wang, ke & Chen, Rui & Yu, Philip. Privacy-Preserving Data Publishing: A Survey of Recent Developments. ACM Comput. Surv, 2010. №4 (14). С. 1-53.
5. Yaseen, Saba & Abbas, Syed & Anjum, Adeel & Saba, Tanzila & Khan, Abid & Malik, Saif & Ahmad, Naveed & Shahzad, Basit & Bashir, Ali. Improved Generalization for Secure Data Publishing. IEEE Access, 2018. С. 27156-27165.
6. Fung, Benjamin & Wang, Ke & Fu, Ada & Yu, Philip. Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques, 2010. 376 с. ISBN: 9780429138737.
7. Li, Ninghui & Li, Tiancheng & Venkatasubramanian, Suresh. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. IEEE 23rd International Conference on Data Engineering (ICDE), 2007. 2. С. 106 - 115.
8. Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. J. Uncertain. Fuzz. Knowl. Sys., 2002. 10 (5), С. 571-588.
9. US Department of Health and Human Services. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) privacy rule, 2014. [Електронний ресурс] / Режим доступу: http://www.hhs.gov/.

10. Simson L. Garfinkel. NISTIR 8053. De-Identification of Personal Information, 2015. [Електронний ресурс] / Режим доступу: http://dx.doi.org/10.6028/NIST.IR.8053
11. Fung B., Wang ke, Wang L., Debbabi M. A framework for privacy-preserving cluster analysis. Conference: Intelligence and Security Informatics, 2008. С. 46 - 51.
12. Emam K., Dankar F. (2008). Protecting Privacy Using k-Anonymity. Journal of the American Medical Informatics Association : JAMIA. 2008. 15(5).
13. 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.
14. 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.
15. Шеннон К. Работы по теории информации и кибернетике. Издательство иностранной литературы. 1963. 830 с.
16. Шнайер, Б. Секреты и ложь. Безопасность данных в цифровом мире. – Пер. с англ. – СПб.: Питер, 2004. с. 432. ISBN: 5-318-00193-9.
17. Быстрый выбор случайных значений из больших таблиц MySQL по условию. [Електронний ресурс] / Доступно: https://habr.com/ru/post/207096/
Дата звернення: Трав. 1, 2022.
18. Greg Robidoux. Retrieving random data from SQL Server with TABLESAMPLE. [Електронний ресурс] / Доступно: https://www.mssqltips.com/sqlservertip/1308/retrieving-random-data-from-sql-server-with-tablesample/.
Дата звернення: Трав. 1, 2022.
19. NOTES ON SQL. [Електронний ресурс] / Доступно: https://sqlrambling/.net/2018/01/24/tablesample-basic-examples.
Дата звернення: Трав. 1, 2022.

REFERECES

1. Dankar, F., Emam, K.E., Neisa, A., Roffey, T.: Estimating the re-identification risk of clinical data sets. BMC Med. Inform. Decis. Mak, 2012. №12 (66). P. 1-15.
2. B.A. Malin, D. Karp, R.H. Scheuermann. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. J. Investig. Med., 2010. 58 (1). P. 11-18.
3. Li, Tiancheng & Li, Ninghui. On the tradeoff between privacy and utility in data publishing. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009. P. 517-526.
4. Fung, Benjamin & Wang, ke & Chen, Rui & Yu, Philip. Privacy-Preserving Data Publishing: A Survey of Recent Developments. ACM Comput. Surv, 2010. №4 (14). P. 1-53.
5. Yaseen, Saba & Abbas, Syed & Anjum, Adeel & Saba, Tanzila & Khan, Abid & Malik, Saif & Ahmad, Naveed & Shahzad, Basit & Bashir, Ali. Improved Generalization for Secure Data Publishing. IEEE Access, 2018. P. 27156-27165.
6. Fung, Benjamin & Wang, Ke & Fu, Ada & Yu, Philip. Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques, 2010. 376 s. ISBN: 9780429138737.
7. Li, Ninghui & Li, Tiancheng & Venkatasubramanian, Suresh. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. IEEE 23rd International Conference on Data Engineering (ICDE), 2007. 2. P. 106 - 115.
8. Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. J. Uncertain. Fuzz. Knowl. Sys., 2002. 10 (5). P. 571-588.
9. US Department of Health and Human Services. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) privacy rule, 2014. available at:http://www.hhs.gov/.
10. Simson L. Garfinkel. NISTIR 8053. De-Identification of Personal Information, 2015. available at: http://dx.doi.org/10.6028/NIST.IR.8053.
11. Fung B., Wang ke, Wang L., Debbabi M. A framework for privacy-preserving cluster analysis. Conference: Intelligence and Security Informatics, 2008. P. 46 - 51.

12. Emam K., Dankar F. (2008). Protecting Privacy Using k-Anonymity. Journal of the American Medical Informatics Association : JAMIA. 2008. 15(5).
13. 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. P. 235-241. ISBN: 978-989-758-474-9.
14. 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. P. 0808-0815. ISBN: 9781665483032.
15. Shennon K. Raboty po teorii informacii i kibernetike. Izdatel'stvo inostrannoj literatury, 1963. 830 s.
16. Shnajer, B. Sekrety i lozh'. Bezopasnost' dannyh v cifrovom mire. – Per. s angl. – SPb.: Piter, 2004. s. 432. ISBN: 5-318-00193-9.
17. Bystryj vybor sluchajnyh znachenij iz bol'shih tablic MySQL po usloviyu. available at: https://habr.com/ru/post/207096/.
Available: May. 1, 2022.
18. Greg Robidoux. Retrieving random data from SQL Server with TABLESAMPLE. available at: https://www.mssqltips.com/sqlservertip/1308/retrieving-random-data-from-sql-server-with-tablesample/.
Available: May. 1, 2022.
19. NOTES ON SQL. available at: https://sqlrambling.net/2018/01/24/tablesample-basic-examples.
Available: May. 1, 2022.

Podoliaka Oksana *PhD of Technical Sciences, docent*
V.N. Karazin Kharkiv National University, Svobody Square 4, Kharkiv, Ukraine, 61022
Podoliaka Oleksii *Senior lecturer*
V.N. Karazin Kharkiv National University, Svobody Square 4, Kharkiv, Ukraine, 61022

Assessing the utility of a public dataset for analytical research

Organizations and agencies release various data intended for analysis, training of artificial intelligence systems, and other research purposes. According to the adopted regulations in the field of personal data protection, public data must be anonymized and protected from various threats of personal data disclosure. Elimination of these threats is realized by reducing the accuracy of data during their preparation for the release. Loss of accuracy obviously leads to a decrease in the usefulness of data for analysis. The paper considers entropy metrics of utility and problems of their computability, as well as metrics of loss of utility of certain subsets of public data.

Objective. To develop effective metrics for assessing the usefulness of a public dataset for analysis, taking into account the requirements of personal data protection.

Research methods. Information security, Shannon's theory of information, Data Governance.

Results. Metrics for assessing information loss and data usefulness for analysis based on the entropy metrics of Shannon's information theory are proposed. Procedures aimed at increasing the speed of calculations of the considered metrics are suggested.

Conclusions. The procedures for building a secure public dataset are described. The application of entropy metrics of Shannon's information theory to assess information loss and data usefulness for analysis is considered. It has been shown that the calculation of these metrics is a complex computational task that is practically impossible for large databases. Procedures aimed at increasing the speed of calculating the considered metrics are proposed. In particular, the creation of a less accurate copy of the original data and the formation of a random sample from a large database to calculate the necessary statistics. The metrics for assessing the usefulness of certain subsets (clusters) of public data are considered in the article.

Keywords: *data privacy, de-identification, data publishing, data utility, GDPR (General Data Protection Regulation).*

Надійшла у першій редакції 04.03.2024, в останній - 01.05.2024.
The first version has been received on 04.03.2024, the final version - on 01.05.2024.