Система запобігання кіберзлочинності у відкритих інформаційних ресурсах на стадії формування контенту

Система предотвращения киберпреступности в открытых информационных ресурсах на стадии формирования контента

Cybercrime prevention system in opened information resources at the stage of content formation

1. **Номер державної реєстрації теми – 0117U004268**

2. **Науковий керівник –** д.т.н., проф. Чертов О.Р., Чертов О.Р., Chertov Oleg R.

3. **Суть розробки, основні результати**

(укр.)

«Публічна інформація у формі відкритих даних» — це публічна інформація у форматі, який дає можливість виконувати її автоматизоване оброблення електронними засобами, вільний та безоплатний доступ до неї, а також її дальше використання. Розпорядники інформації зобов'язані надавати публічну інформацію у формі відкритих даних на запит, оприлюднювати й регулярно оновлювати її на державному веб-порталі відкритих даних та на своїх веб-сайтах. Відповідна інформація є дозволеною для її дальшого вільного копіювання, опублікування, використання та розповсюдження.

За таких обставин потенційні зловмисники за допомогою методів аналізу даних можуть віднайти приховані закономірності, структури розподілу даних, виявити й мати безпосередній доступ до інформації з обмеженим доступом, якщо така міститиметься у відкритих даних, зокрема й у неявному вигляді. Навіть вилучення, перед оприлюдненням даних, відповідних атрибутів не гарантує схоронності інформації. Методи забезпечення групової анонімності даних дають змогу захистити інформацію про групи осіб, наприклад, інформацію про територіальний та інші типи розподілів.

Зважаючи на велику кількість відкритих інформаційних ресурсів в Україні, а саме: сайти юридичних і фізичних осіб, бази даних, державні й корпоративні реєстри, інформаційні сховища та інформаційні колектори, ресурси громадських організацій і соціальних мереж тощо, у яких розміщують відкриті дані, що можуть містити інформацію з обмеженим доступом, нагальною є потреба забезпечення захисту такої інформації. Застосування для цього методів і засобів технічного захисту інформації є просто неможливим, оскільки дані є відкритими й доступними для широкого кола користувачів.

Таким чином, актуальною і доцільною є задача розроблення моделей, методів та інформаційних технологій, що дозволять створити систему запобігання кіберзлочинності у відкритих інформаційних ресурсах.

Розроблені математичні моделі забезпечення індивідуальної та групової анонімності мікрофайлів (для даних, що можуть містити інформацію з обмеженим доступом) та методи забезпечення анонімності даних, які не передбачають участі експерта в оцінюванні якості одержуваних результатів.

Також розроблено структурне представлення та архітектуру інформаційної технології запобігання кіберзлочинності у відкритих інформаційних ресурсах на стадії формування контенту. У відповідності до Державних стандартів розроблене організаційне, інформаційне, математичне та технічне забезпечення інформаційної технології запобігання кіберзлочинності у відкритих інформаційних ресурсах на стадії формування контенту із застосуванням методів індивідуальної та групової анонімізації на рівні абсолютних та відносних даних.

Застосування цієї інформаційної технології дасть змогу значно розширити кількість наборів даних, доступних для публічного дослідження, що, у свою чергу, дасть змогу суттєво зменшити суб'єктивізм під час планування соціально-економічного розвитку країни.

(рос.)

«Публичная информация в форме открытых данных» – это публичная информация в формате, который дает возможность выполнять ее автоматизированную обработку электронными средствами, свободный и бесплатный доступ к ней, а также ее дальнейшее использование. Распорядители информации обязаны предоставлять публичную информацию в форме открытых данных на запрос, обнародовать и регулярно обновлять ее на государственном веб-портале открытых данных и на своих веб-сайтах. Соответствующая информация является разрешенной для ее дальнейшего свободного копирования, опубликования, использования и распространения.

При таких обстоятельствах потенциальные злоумышленники с помощью методов анализа данных могут обнаружить скрытые закономерности, структуры распределения данных, выявить и получить непосредственный доступ к информации с ограниченным доступом, если такая будет содержаться в открытых данных, в частности и в неявном виде. Даже удаление, перед обнародованием данных, соответствующих атрибутов не гарантирует сохранности информации. Методы обеспечения групповой анонимности данных дают возможность защитить информацию о группах лиц, к примеру, информацию о территориальном и других типах распределений.

Учитывая большое количество открытых информационных ресурсов в Украине, а именно: сайты юридических и физических лиц, базы данных, государственные и корпоративные реестры, информационные хранилища и информационные коллекторы, ресурсы общественных организаций и социальных сетей и т.п., в которых размещают открытые данные, которые могут содержать информацию с ограниченным доступом, насущной является потребность обеспечения защиты такой информации. Применение для этого методов и средств технической защиты информации просто невозможно, поскольку данные являются открытыми и доступными для широкого круга пользователей.

Таким образом, актуальной и целесообразной является задача разработки моделей, методов и информационных технологий, которые позволят создать систему предотвращения киберпреступности в открытых информационных ресурсах.

Разработаны математические модели обеспечения индивидуальной и групповой анонимности микрофайлов (для данных, которые могут содержать информацию с ограниченным доступом) и методы обеспечения анонимности данных, которые не предусматривают участия эксперта в оценивании качества получаемых результатов.

Также разработаны структурное представление и архитектура информационной технологии предотвращения киберпреступности в открытых информационных ресурсах на стадии формирования контента. В соответствии с Государственными стандартами разработано организационное, информационное, математическое и техническое обеспечение информационной технологии предотвращения киберпреступности в открытых информационных ресурсах на стадии формирования контента с применением методов индивидуальной и групповой анонимизации на уровне абсолютных и относительных данных.

Применение этой информационной технологии позволит значительно расширить количество наборов данных, доступных для публичного исследования, что, в свою очередь, позволит существенно уменьшить субъективизм во время планирования социально-экономического развития страны.

(англ.)

“Public information in the form of open data” is public information in a format that makes it possible to carry out its automated processing by electronic means, open and free access to it, as well as its further using. Information managers are required to provide public information in the form of open data upon request, to publish and regularly update it on a state open data web portal and on their websites. Such information is permitted for its further free copying, publication, use and distribution.

In such circumstances, potential attackers using data analysis methods can detect hidden patterns, data distribution structures, identify and gain direct access to information with limited access, if such data is contained in open data, even in an implicit form. Even removing the corresponding attributes before the publication of data does not guarantee the safety of information. Methods of ensuring group data anonymity make it possible to protect information about groups of people, for example, information about territorial and other types of distributions.

Taking into account the large number of open information resources in Ukraine, namely: websites of legal entities and individuals, databases, state and corporate registries, information repositories and information collectors, resources of public organizations and social networks, etc., in which open data is posted, which may contain information with limited access, the task of ensuring the protection of such information is an urgent need. The application of methods and means of technical protection of information for such task is simply impossible, since the data is open and accessible to a wide range of users.

Thus, the urgent and appropriate is the task of developing models, methods and information technologies that will create a system for preventing cybercrime in open information resources.

Mathematical models have been developed to ensure individual and group anonymity of microfiles (for data that may contain information with limited access) and methods to ensure anonymity of data that do not provide for expert participation in assessing the quality of the results obtained.

The structural representation and architecture of information technology for preventing cybercrime in open information resources at the stage of content formation was also developed. In accordance with State standards, the organizational, informational, mathematical and technical support of information technology for preventing cybercrime in open information resources at the stage of content formation using individual and group anonymization methods at the level of absolute and relative data has been developed.

The use of this information technology will significantly expand the number of data sets available for public research, which, in turn, will significantly reduce subjectivity during the planning of socio-economic development of the country.

4. Наявність охоронних документів на об’єкти права інтелектуальної власності

Не подавались на реєстрацію.

5. Порівняння зі світовими аналогами

У класичних та сучасних роботах в області «збереження приватності при публікації даних» (privacy-preserving data publishing) досліджуються лише питання забезпечення індивідуальної анонімності даних. Значна частина відповідних методів реалізована в пакеті μ-Argus, який де-факто став стандартом у галузі і зараз вільно розповсюджується. У деяких випадках також застосовують методи, відмінні від методів маскування, наприклад, методи генерації синтетичних даних чи комбіновані методи, що використовують і вихідні, і синтетичні дані. Для підготовки мікрофайлів з даними респондентів такі методи не прийнятні, оскільки, не знаючи цілей такого аналізу, неможливо згенерувати адекватні набори синтетичних даних.

На сьогоднішній день світовий пріоритет у розробленні методів забезпечення саме групової анонімності даних належить колективу авторів проекту та його науковому керівнику О. Р. Чертову.

Уперше задачу забезпечення групової анонімності сформульовано як задачу пошуку максимального потоку мінімальної вартості, у якій на архітектуру мережі накладено нечіткі обмеження, і запропоновано оригінальний метод її розв’язання на основі міметичних обчислень, який полягає у формуванні відповідних нечітких обмежень та дальшому застосуванні гібридного еволюційного алгоритму, де вони враховуються у функції пристосованості, що дає можливість одержувати розв’язки цієї задачі допустимої вартості, тобто з прийнятним ступенем сумісні з накладеними обмеженнями.

Удосконалено метод виявлення підгруп, який відрізняється від існуючих новою мірою якості нечітких правил для опису підгруп, яка враховує непропорційно велику відносну перевагу кількостей елементів підгрупи над кількостями елементів поза нею в окремих областях простору ознак, що дає змогу виділяти підгрупи малого обсягу та високої локальної концентрації.

Удосконалено моделі груп респондентів, що використовуються для порушення анонімності даних про ці групи, які відрізняються від існуючих врахуванням базових атрибутів мікрофайлу, що дає змогу адаптувати задачу забезпечення групової анонімності до груп респондентів, анонімність яких можна порушити у випадку вилучення з мікрофайлу сутнісних атрибутів.

Розроблено та апробовано інформаційну технологію генерації рекомендацій на основі даних із попередньо захищеною приватністю.

Розроблено інформаційну технологію забезпечення групової анонімності даних, яка відрізняється від існуючих засобів забезпечення анонімності тим, що враховує комбінації значень базових атрибутів мікрофайлу, що дає можливість приховувати особливості розподілу інформації про групи записів, відносно яких існує ризик порушення анонімності у випадку вилучення з мікрофайлу сутнісних атрибутів, забезпечуючи при цьому прийнятний рівень спотворення даних.

6. Економічна привабливість для просування на ринок

Кошти державного бюджету використано для розроблення методів, інформаційної технології та системи запобігання кіберзлочинності ще на стадії формування контенту у відкритих інформаційних ресурсах організаційних систем державної та приватної форм власності й окремих фізичних осіб. Відповідні захищені мікрофайли потенційно є важливими для владних структур центрального та місцевого рівня для розрахунку видатків із державного бюджету на соціальні потреби населення країни і визначення міжбюджетних трансфертів у галузях охорони здоров'я, соціального захисту населення, культури, міського громадського транспорту, для генерації рекомендацій стосовно вкладення муніципальних та державних коштів для стимулювання певних дій населення (збільшення народжуваності, закріплення на місцевості проживання, покращення енергозбереженості тощо). Однак ці дані можуть містити інформацію з обмеженим доступом, яку не можна захистити традиційними методами і засобами технічного захисту інформації у відповідності до чинного законодавства. Наразі отримано лист про наміри про співпрацю від потенційного замовника – товариства з обмеженою відповідальністю «ГРІН ЛАЙТ КОРПОРАТИВНІ РІШЕННЯ», яке згідно з Договором № 107 від 17.12.2018 р. між цим товариством та Державною службою статистики України є розробником програмно-апаратного комплексу автоматизованої системи для збору та оброблення даних пробного перепису населення (АС ППН). Захищені проблемно-орієнтовані мікрофайли за даними Всеукраїнського перепису населення за кожним регіоном України підготовлені та передані Державній службі статистики України та неприбутковій організації Informed Decisions (США) при Науковому Товаристві імені Тараса Шевченка в Нью-Йорку.

7. Потенційні користувачі (галузі, міністерства, підприємства, організації)

Система запобігання кіберзлочинності у відкритих інформаційних ресурсах на стадії формування контенту матиме застосування у органах державної влади, підприємств державної та приватної форм власності, що працюють з даними, призначеними для розміщення у відкритих, публічних інформаційних ресурсах.

Систему запобігання кіберзлочинності у відкритих інформаційних ресурсах на стадії формування контенту може бути передано територіальним органам Державної служби статистики України, міжнародним статистичним організаціям, органам державної влади й місцевого самоврядування, підприємствам, установам та організаціям корпоративного сектору економіки, установам Національної академії наук України, окремим громадським організаціям тощо.

8. Стан готовності розробки

Систематизовано результати аналізу існуючих рішень методів та моделей забезпечення індивідуальної та групової анонімності. Розроблені математичні моделі та методи забезпечення індивідуальної та групової анонімності мікрофайлів. Створена архітектура інформаційної технології та її організаційне, математичне, інформаційне та технічне забезпечення. Розроблена інформаційна технологія генерації рекомендацій на основі даних із попередньо захищеною приватністю. Проведено захист мікрофайлів за даними Всеукраїнського перепису населення. Створено експериментальний зразок автоматизованої системи запобігання кіберзлочинності у відкритих інформаційних ресурсах на стадії формування контенту.

9. Існуючі результати впровадження

Результати роботи впроваджено у навчальний процес кафедри прикладної математики у вигляді лабораторної роботи «Проект інженерії системи запобігання кіберзлочинності у відкритих інформаційних ресурсах на стадії формування контенту для ВНЗ України» в рамках курсу «Управління проектами». Підготовлені за даними Всеукраїнського перепису населення мікродані застосовуються у Державній службі статистики України та неприбутковій організації Informed Decisions (США) при Науковому Товаристві імені Тараса Шевченка в Нью-Йорку.

10. Назва організації, телефон, E-mail

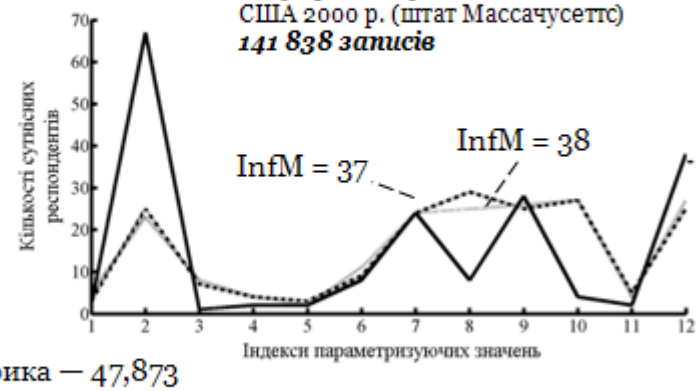
КПІ ім. Ігоря Сікорського, кафедра прикладної математики, (044) 236-95-05, pmakpi@gmail.com

11. Фото розробки - Слайди

Приклад застосування

3000 особин:
— 2 693 розв'язки

Мікрофайл: перепис населення США 2000 р. (штат Массачусетс)
141 838 записів



Середня метрика — 47,873

Забезпечення анонімності досягається за рахунок зміни не більше, ніж $\frac{47,783}{7 \cdot 141\ 838} \approx 0,005\%$ значень атрибутів

Експеримент

Мікрофайл: перепис населення США 2000 р. (штат Флорида)
334 364 записів

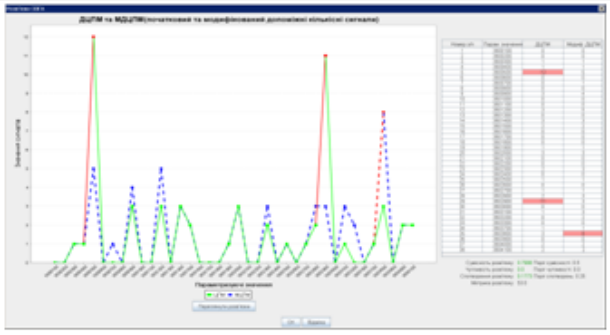
1 регіон — база Пенсакола
2 регіон — база Еґлін
5 регіон — база Джексонвіл

Матриця невідповідностей

$$\mathbf{Z} = \begin{pmatrix} 2 & 1 \\ 0 & 15 \end{pmatrix}$$

Точність класифікації — 0,944
Статистика *J* Йовдена — 0,938

Експеримент (модель на основі сторонніх даних)



Середня метрика – 62,518

Забезпечення анонімності досягається за рахунок зміни значень атрибутів $\frac{62,518}{13 \cdot 1380\,924} \approx 0,0003\%$

Розв'язання задачі силами колективу з 5 фахівців – **8 год 20 хв.**, за допомогою описаної в літературі ІТ – **20 год 5 хв.** (у **2,4 разу** довше)

Архітектура інформаційної системи для реалізації ІТ

12. Перелік публікацій за матеріалами досліджень за період виконання розробки

1. *Chertov, O., Tavrov, D. Improving efficiency of providing data group anonymity by automating data modification quality evaluation. Eastern-European Journal of Enterprise Technologies. - 5/4 (89) 2017, P. 31-39.*
2. *Aleksandrova, M., Brun, A., Boyer, A., Chertov, O. Identifying representative users in matrix factorization-based recommender systems: application to solving the content-less new item cold-start problem. Journal of Intelligent Information Systems, 2017, 48 (2), P. 365-397.*
3. *Aleksandrova, M., Brun, A., Chertov, O., Boyer, A. Sets of contrasting rules: A supervised descriptive rule induction pattern for identification of trigger factors. Proceedings - IEEE 28th International Conference on Tools with Artificial Intelligence. - 2017. P. 431-435.*
4. *Chertov, O., Malchykov, V. Determining Optimal Dilation Factor of Non-Dyadic Wavelet Transform. 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON). May 29 2017-June 2 2017. P. 297-300.*
5. *Чертов О.Р., Тавров Д.Ю. Забезпечення групової анонімності як складова CSID-процесу обробки даних. Штучний інтелект, 2017, № 3-4, С. 127-132.*

6. *Aleksandrova, M., Chertov, O., Brun, A., Boyer, A. Contrast classification rules for mining local differences in medical data. 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2017, v. 2, IEEE, P. 880-883.*
7. *Chertov, O., Tavrov, D. Improving efficiency for ensuring data group anonymity by developing an information technology. Eastern-European Journal of Enterprise Technologies. - 6/4 (96) 2018, P. 41-56.*
8. *Chertov, O., Rudnyk, T., Palchenko, O. Search of phony accounts on Facebook: Ukrainian case. Proc. 2018 International Conference on Military Communications and Information Systems, ICMCIS 2018, 22-23 May 2018. 1-4 p.*
9. *Pavlenko P., Tavrov D., Temnikov V., Zavgorodniy S., Temnikov A. The Method of Expert Evaluation of Airports Aviation Security Using Perceptual Calculations. The 9th IEEE International Conference on Dependable Systems, Services and Technologies, DESSERT'2018 24-27 May, 2018, Kyiv, Ukraine, P. 32-35.*
10. *Pavlov D., Chertov O. How Click-Fraud Shapes Traffic: A Case Study. In: Chertov O., Mylovanov T., Kondratenko Y., Kacprzyk J., Kreinovich V., Stefanuk V. (eds) Recent Developments in Data Science and Intelligent Analysis of Information. ICDSIAI 2018. Advances in Intelligent Systems and Computing, Springer, Cham, vol. 836, pp. 238-248.*
11. *Sokhatskyi M., Maslianko P. Constructive Proofs of Heterogeneous Equalities in Cubical Type Theory. In: Chertov O., Mylovanov T., Kondratenko Y., Kacprzyk J., Kreinovich V., Stefanuk V. (eds) Recent Developments in Data Science and Intelligent Analysis of Information. ICDSIAI 2018. Advances in Intelligent Systems and Computing, Springer, Cham, vol. 836, pp. 305-318.*
12. *Wiktorski T., Demchenko Y., Chertov O. Data Science Model Implementation for Various Types of Big Data Infrastructures, Proc. 5th IEEE STC CC Workshop on Methods in Cloud Computing, Big Data, and Data Science (DTW2019), part of the eScience 2019 Conference, September 24–27, 2019, San Diego, California, USA.*
13. *Marharyta Aleksandrova. Matrix Factorization and Contrast Analysis Techniques for Recommendation; дисертація захищена 7 липня 2017 р. в Університеті Лотарингії (Франція) в рамках навчання в подвійній аспірантурі: http://docnum.univ-lorraine.fr/public/DDOC_T_2017_0080_ALEKSANDROVA.pdf.*

13. Надати ключові слова до розробки: кіберзлочинність, групова анонімність, відкриті дані, еволюційні алгоритми.