УДК 343.9
DOI https://doi.org/10.32782/2524-0374/2022-5/136

ОЧИСТКА ДАНИХ ЯК ОДИН ІЗ МЕТОДІВ ОБРОБКИ ДАНИХ ДЛЯ КРИМІНОЛОГІЧНИХ ДОСЛІДЖЕНЬ

DATA CLEANING AS ONE OF THE METHODS OF DATA PROCESSING FOR CRIMINAL INVESTIGATIONS

Коростельова Л.А., ад'юнктка кафедри кримінально-правових дисциплін
Луганський державний університет внутрішніх справ імені Е.О. Дідоренка

У статті розглянуто один із інноваційних методів обробки даних *Data cleaning*. Визначені проблемні питання, що виникають у кримінології під час обробки даних для досліджень. Надано поняття очищення даних і його використання в контексті науки про дані, зокрема і в кримінологічній науці. Описано етапи процесу очищення даних. Визначено, що очищення даних вважається основоположним елементом основ науки про дані, оскільки відіграє важливу роль в аналітичному процесі та пошуку надійних відповідей.

Запропоновано та розглянуто концепції очищення даних для кримінологічних досліджень. Проведено аналіз методів і технологій очищення даних на кожному із етапів процесу з врахуванням його особливостей для кримінологічної науки. Побудована процедура очищення даних для систематизації методів у реалізації моделі для кримінологічного дослідження. Якість персональних даних є проблемою, що значно знижує результативність аналізу.

Визначена авторська думка, щодо застосування спеціалізованих інструментів і методів, що дають змогу перетворити зібрані «сирі» дані у цінну інформацію, що використовується в процесі кримінологічної розвідки.

Обґрунтовано висновок, що метод очищення даних (*Data cleaning)* для кримінологічної науки, і зокрема для кримінологічних досліджень має перспективний напрям. У висновках обґрунтовано запропоновані зміни для вдосконалення кримінологічної науки. А також визначено і сформовано критерії підбору алгоритмів очищення даних від випадково виникаючих помилок в процесі отримання даних. Рекомендовано використання шаблонів обробки на основі сформованих метаданих вхідного потоку для використання вже розрахованих алгоритмів очищення. У подальшій роботі буде розглянуто більш ефективне формування метаданих, їх формат та зберігання. У статті доведено, що інноваційний метод очищення даних (*Data cleaning)* потребує впровадження і подальшого розвитку у кримінологічній науці, зокрема і для кримінологічних досліджень, а також для перепідготовки криміналогів на новий рівень розвитку.

Ключові слова: *Data cleaning,* методи кримінології, кримінологічні дослідження, наука про дані, метадані.

The article considers one of the innovative data processing methods – Data cleaning. The problematic issues that arise in criminology during data processing for research are identified. The concept of data purification and its use in the context of data science, in particular in criminology, is given. The stages of the data cleaning process are described. It is determined that data cleaning is considered a fundamental element of the foundations of data science, as it plays an important role in the analytical process and the search for reliable answers.

Concepts of data cleaning for criminal investigations are offered and considered. The analysis of methods and technologies of data cleaning at each stage of the process and its features for criminology is carried out. The procedure of data cleaning for the systematization of methods in the realization of a model for criminal investigations is constructed. The quality of personal data is a problem that significantly reduces the effectiveness of the analysis.

The author's opinion on the use of specialized tools and methods that allow transforming the collected "raw" data into valuable information used in the process of criminal intelligence is defined.

The conclusion that the method of data cleaning for criminology, and in particular for criminal investigations has a promising direction, is substantiated. The conclusions substantiate the proposed changes to improve criminology as a science. Also, the criteria for selecting algorithms for cleaning data from accidental errors in the process of obtaining data are defined and formulated. It is recommended to use processing templates based on the generated input data metadata to use the already calculated cleaning algorithms. Further work will involve considering of more efficient formation of metadata, their format, and storage. The article proves that the innovative method of data cleaning requires the introduction and further development in criminology, in particular for criminal investigations, as well as for the retraining of criminologists for a new level of development.

Key words: data cleaning, methods of criminology, criminal investigations, data science, metadata.

Постановка проблеми. Історичний досвід запобігання злочинності вказує на важливість застосування технологічних інновацій для підвищення рівня ефективності цієї діяльності. Це має бути рушійною силою, що веде до реформування стратегій запобігання злочинності із залученням не тільки правоохоронних органів але й громадськості. Сучасною тенденцією щодо запобігання злочинності є використання можливостей мережі Інтернет, зокрема інформаційних інтернет-технологій, що постійно розвиваються і відкривають нові можливості в різних сферах життя суспільства. Кримінологічні дослідження останнього десятиліття великою мірою присвячені саме вказаним проблемам [1, с. 243].

Дійсно, розвиток технологій значно вплинув на суспільство і на злочинність. В сучасному світі все більше виникає потреб у дослідженні злочинності із використанням сучасних автоматизованих методів збору, обробки і аналізу інформації.

Розвиток методів запису і зберігання даних викликав бурхливе зростання об'ємів збираної і аналізованої інформації. Об'єми даних настільки значні, що людина просто не спроможна проаналізувати їх самостійно, хоча необхідність проведення такого аналізу цілком очевидна, адже в цих «сирих даних» закладено знання, які можуть бути використані при ухваленні рішень [2, с. 97].

Аналіз останніх досліджень і публікацій. Окремі аспекти досліджуваної проблеми розглядались у працях вітчизняних учених, зокрема, Головкіна Б, В. В. Сташиса, С. А. Тарарухіна, В. П. Тихого, І. К. Туркевич, П. Л. Фріса, В. І. Шакуна, С. С. Яценка та інших.

Метою статті є впровадження методу очистки даних в кримінологічні дослідження для отримання якісних кримінологічних розвідок, що вплинуть на протидію злочинності.

Виклад основного матеріалу. Стійкий прогрес розвитку комп'ютерних інформаційних технологій минулих трьох десятиліть призвів до появи великої кількості потужних комп'ютерів, програмного та апаратного забезпечення для зберігання та обробки даних. Ці технології зробили доступними користувачам величезну кількість баз даних та інших сховищ інформації для пошуку та вилучення з них інформації, а також для аналізу даних. Завдяки цьому розвитку новітніх технологій WEB простір став найбільшим з відкритих джерел інформації сучасності.

За статистикою 93% нової інформації світу зберігається в електронному вигляді і є в тій чи іншій мірі доступною користувачеві. Окрім цього інформація у Web охоплює майже всі можливі теми і існує майже у всіх доступних формах (таблиці, текст, графічна інформація, відео, звук). Вона є динамічною і постійно змінюється. Зі зростанням об'ємів інформації у Web зростає і необхідність розвитку та вдосконалення засобів вилучення і обробки інформації вебсистеми [3, с. 37].

Одним із сучасних методів обробки інформаціє є метод очистки даних (*Data cleaning*).

Очистка даних- виявлення та видалення помилок і не відповідностей даних з метою підвищення якості даних. Проблеми з якістю даних відображаються в єдиній колекції даних, таких як файли та бази даних, наприклад, із-за неправильного опису при введенні даних, відсутній інформації чи інші незрозумілі дані. Коли необхідно інтегрувати кілька джерел даних, наприклад, в зберіганні даних, потрібно об'єднати системи баз даних або глобальні інформаційні вебсистеми [4].

Потреба в методі очищення даних зосереджена на покращенні якості даних наукових розвідок, за рахунок функції автоматизованої очистки даних.

Так, в свою чергу Редман припустив, що якщо використовувати зазначений метод у наукових дослідженнях, то слід очікувати коефіцієнт польової помилки 1–5% [5].

Варто зазначити, що в кримінологічній науці завжди існували проблемні питання щодо якості отримання даних та їх покращення.

В сучасних реаліях все більше виникає потреба у використанні методології науки про дані, зокрема застосовуючи методи цієї методології. Автоматизований процес очищення даних може включати видалення помилок, перевірку даних та покращення даних. Для детального розгляду використання методу очистки даних, ми повинні визначити критерії якості даних, які включають: точність, повноту, послідовність і однорідність.

В більшості випадків для початку процесу очистки даних необхідно визначити основні стовбці, які відносяться до певного типу даних, це необхідно для того, щоб перейти до наступної операції із методом типу даних «object». Типи даних «object» зустрічаються у декількох стовбцях і їх необхідно корегувати у всіх стовбцях до правильних типів даних, зокрема використовуючи критерії методу очистки даних.

Загалом існує декілька критеріїв очистки даних:

1. Видалення небажаних спостережень. Оскільки одна з основних цілей очищення даних полягає в тому, щоб переконатися, що набір даних не містить небажаних спостережень, це класифікується як перший крок до очищення даних. Небажані спостереження в наборі даних бувають 2 типів, а саме; дублікати та невідповідності.

2. Повторювані спостереження. Зазвичай це виникає, коли набір даних створюється в результаті об'єднання даних з двох або більше джерел. Це також може відбуватися в деяких інших випадках, зокрема, коли респондент робить кілька запитів на опитування або помиляється під час введення даних.

3. Нерелевантні спостереження. Невідповідні спостереження – це ті, які насправді не відповідають конкретній проблемі, яку ви намагаєтеся вирішити.

4. Виправлення структури даних. Після видалення небажаних спостережень наступне, що потрібно зробити, це переконатися, що потрібне спостереження добре структуровано. Під час передачі даних можуть виникнути структурні помилки через незначну людську помилку або некомпетентність персоналу, що вводить дані.

5. Відфільтрування даних. Щоб підвищити продуктивність наукової розвідки необхідно провести етап фільтрування даних. Онова яких полягає у видаленні відповідних точок даних, які відрізняються від інших спостережень у наборі даних [6].

Такий послідовний критерій обробки даних для криміологічних досліджень виключає людський фактор помилок і тим самим оптимізує роботу самого дослідження.

Окрім критеріїв очистки даних існує необхідність зазначити основні етапи зазначеного методу.

1. Аналіз даних – виявлення видів помилок і невідповідностей, що підлягають видаленню. Поряд з ручною перевіркою даних або їхніх шаблонів, треба використовувати аналітичні програми для отримання метаданих про властивості даних і виявлення проблем якості даних.

2. Визначення послідовності і правил перетворення даних. Залежно від кількості джерел даних, ступеня їхньої неоднорідності та забрудненості даних, вони можуть вимагати достатньо широкого перетворення та очищення. Іноді для відображення джерел для загальної моделі даних використовується трансляція схеми; для сховищ даних, зазвичай, використовується реляційне зображення. Перші кроки з очищення даних можуть скоригувати проблеми окремих джерел даних і підготувати дані для інтеграції. Подальші кроки спрямовані на інтеграцію схеми/даних та усунення проблем множинності елементів, наприклад, дублікатів. Для сховищ даних у процесі ETL (Extract Transform, Load – «видобування, перетворення, завантаження») визначаються методи контролю і потік даних, що підлягає перетворенню та очищенню. Перетворення даних, що пов'язане зі схемою, визначається за допомогою мови декларативного запиту (маппірування, Mapping Composition), забезпечуючи, у такий спосіб, автоматичну генерацію коду перетворення. У процесі перетворення має бути можливість запуску написаного користувачем коду очищення і спеціальних засобів. Етапи перетворення можуть вимагати зворотного зв'язку з користувачем по тих елементах даних, для яких відсутня вбудована логіка очищення.

3. Підтвердження – правильність і ефективність процесу і визначення перетворення. Це здійснюється шляхом тестування та оцінювання. Під час аналізу, проектування та підтвердження може знадобитися безліч ітерацій, наприклад, з огляду на те, що деякі помилки стають помітні тільки після певних перетворень.

4. Перетворення – виконання перетворень або в процесі ETL для завантаження і оновлення сховища даних, або при відповіді на запити з множини джерел. Процес перетворення вимагає великих обсягів метаданих – наприклад, схем, характеристик даних рівня схеми, означень технологічного процесу тощо. Для узгодженості, гнучкості та спрощення використання в інших випадках, ці метадані повинні зберігати в репозиторії на основі СУБД. Для підтримки якості даних ґрунтовна інформація про процес перетворення має записуватися і в репозиторій, і в трансформовані елементи даних, особливо інформація про повноту та актуальність початкових даних і походження інформації про першоджерело трансформованих об'єктів та проведені з ними зміни.

5. Протитечія очищених даних – заміна забруднених даних у першоджерелах на очищені. Після того, як помилки (окремого джерела) видалені, очищені дані мають замістити забруднені дані в початкових джерелах, щоб покращені дані потрапили і в успадковані застосування і надалі при витяганні не вимагали додаткового очищення. Для сховищ даних очищені дані містяться в області зберігання даних.

6. Попереднє опрацювання даних – комплекс методів і алгоритмів, які застосовуються в аналітичному додатку з метою підготовки даних до виконання конкретного завдання і приведення їх у відповідність до вимог, що обумовлені специфікою завдання і способами його виконання [7, с. 241].

Але в той же час необхідно зазначити основні недоліки методу очистки даних в кримінологічних дослідженнях.

На нашу думку під час обробки даних кримінологи можуть втратити корисну інформацію через неповні дані. Це дуже часто зустрічається у випадках, коли відсутні спостереження.

По-друге, деякі автоматизовані інструменти очищення даних не дуже розумні і можуть призвести до неправильної обробки деяких спостережень у наборі даних.

По-третє, очищення даних може зайняти багато часу, особливо при роботі з великими даними.

Варто також зазначити, що метод очищення даних має задовольняти низку критеріїв. По-перше, він повинен виявляти і видаляти всі основні помилки і невідповідності, і в окремих джерелах даних, і при інтеграції декількох джерел. По-друге, метод повинен підтримуватися певними інструментами, щоб скоротити обсяги ручної перевірки та програмування, і бути гнучким у роботі з додатковими джерелами. Очищення даних не проводиться незалежно від пов'язаних зі схемою перетворення даних, що виконуються на основі складних метаданих. Інфраструктура технологічного процесу має особливо підтримуватися для сховищ даних, забезпечуючи ефективне і надійне виконання всіх етапів перетворення даних для множини джерел і великих наборів даних [7, с. 244].

Висновки. Отже, підсумовуючи вище зазначене, можна зробити висновок, що із зростанням і цифротизацією за останні 20 років- дані стали одними із найцінніших речей у світі. У світі зростає кількість користувачів соціальних мереж, пошукових систем, вебсайтів тощо. Однак проблема, з якою багато із нас стикається, полягає у тому, що більшість даних або неправильні, або повні невідповідності. Метод очищення даних є одним із найважливіших кроків на шляху до досягнення якості даних для кримінологічних досліджень. Сфера використання методу очистки даних має перспективний напрям у кримінологічній науці.

Але використання такого міждисциплінарного підходу вимагає об'єднання зусиль українських фахівців із правої та IT-сфери для створення багатьох унікальних алгоритмів обробки інформації, щоб створити сучасну кримінологічну науку з широкими можливостями.

ЛІТЕРАТУРА

1. Бугера О. Генезис та сучасний стан використання мережі Інтернет для запобігання злочинності. *Підприємство, господарство і право.* 2018. Вип. 9. С. 243–246.

2. Петренко А. Grid та інтелектуальна обробка даних Data Mining. *Системні дослідження та інформаційні технології.* 2008. № 4. С. 97–110.

3. Бурматова М., Оленін М.В. Аналіз вимог до автоматизованих методів вилучення даних про однотипні обєкти з WEB-простору. *Інженерія програмного забезпечення.* 2010. Вип. 2. С. 37–45.

4. Redman T. Data Quality: The Field Guide. *MA: Digital Press.* 2001. URL: https://dl.acm.org/doi/book/10.5555/362427 (дата звернення 22.03.2022)

5. Erhard Rahm, Hong Hai Do Data Cleaning: Problems and Current Approaches. URL: https://habr.com/ru/post/548220/ (дата звернення: 24.03.2022).

6. Formplus Blog. Data Cleaning: 7 Techniques + Steps to Cleanse Data. URL: https://www.formpl.us/blog/data-cleaning (дата звернення: 24.03.2022).

7. Верес О. Онтологія очищення даних. *Вісник Національного університету «Львівська політехніка».* 2015. № 814. С. 237–245.