УДК 621.382

ОЧИЩЕННЯ ДАНИХ ДЛЯ ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ ОБРОБКИ ДАНИХ В МЕРЕЖАХ ІОТ

Гребініченко М.В.
Інститут телекомунікаційних систем КПІ ім. Ігоря Сікорського, Україна
Email: grebinichenko.maria@gmail.com

DATA CLEANSING FOR INCREASING PERFORMANCE IN IOT NETWORKS

In this article is studying problem of preprocessing big amount of data in IoT networks. One of the methods for resolving this problem is developing new data cleansing dynamic algorithm. The benefits of application of this method are presented and described.

У даній статті розглядається проблема попередньої обробки великих даних у мережах IoT. Один з методів вирішення даної задачі - розробка нового динамічного алгоритму очищення даних. Переваги застосування даного методу також представлені.

Зі стрімким поширенням застосування розумних систем Internet Of Things, зросла і кількість інформації, що підлягає обробці на сервері, а також і кількість факторів, що можуть вносити похибки до датасетів або призводити до часткової втрати даних. Тому з’явилася необхідність підвищення достовірності кінцевого результату роботи мережі IoT за рахунок очищення даних від випадково виникаючих помилок.

На вхідні дані, які пізніше піддаються обробці та аналізу, впливають ряд різноманітних чинників, як збирання та формування набору з різних джерел, передача їх через світову мережу, умови збору інформації з датчиків (температурний, механічний впливи), програмний збій та безліч інших. Кожен з цих факторів потенційно знижує якість даних, що підлягають обробці. Як наслідок у кінцевих обрахунках можуть виникати невалідні та хибні значення, що практично унеможливлює отримання корисних результатів із наборів великих даних.

Існує ряд алгоритмів та методів, що розв‘язують означену проблему, проте вони не є максимально ефективними, чи розв‘язують проблему очищення даних частково. Необхідна модифікація алгоритму попередньої обробки даних для забезпечення гарантовано більш точного результату із мінімальним застосуванням часових та розрахункових ресурсів.

Головною ідеєю виконуваного дослідження є визначення такого алгоритму очищення даних, що дозволив би гарантовано підвищити достовірність кінцевого результату обробки даних із максимальною ефективністю.

Проте не всі дані можуть бути зібрані з різних джерел, а отже містити різні структури моделей та мати різні зв‘язки між параметрами, або містити пропущені дані, бо датчики збору інформації відпрацювали високоякісно протягом всього періоду роботи. Для кожного окремого випадку та кожного набору даних, що йде на обробку, має бути встановлений відповідний процес очистки.

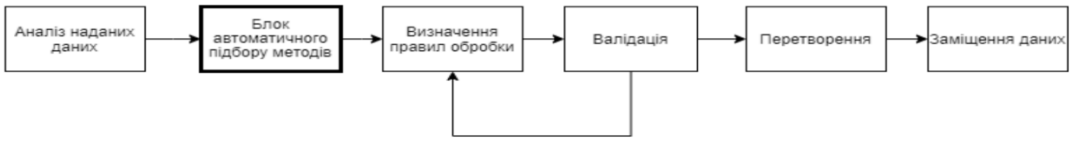

Рис. 1. Блок-схема модифікованого процесу очистки даних.

Запропонований додатковий блок обробки даних має на меті проаналізувати вхідний датасет та адаптивно створити унікальний порядок дій щодо очистки. Для вибору алгоритмів для обробки мають бути сформовані чіткі критерії вибору. Таким чином вдасться розробити алгоритм, що динамічно буде здатний скласти порядок та складові обробки, і тим самим не лише провести ефективну очистку даних, а й знизити навантаження на систему та кількість часових та програмних ресурсів.

Таблиця 1. Компоненти для динамічної попередньої обробки даних.

Компонент очищення	Тип даних	Короткий опис роботи	Умови автоматичного запуску
Фільтрування даних	Числовий	Видалення нерелевантних пікових значень. Після утворення схеми об'єкту даних, користувачу пропонується для кожного ввести максимальне та мінімальне значення.	Відсоток записів , що не є релевантними за початковими умовами більший за Z_0
Відновлення пропущених значень	Числовий	Заповнення відсутніх значень параметрів у наборах даних. Змінна заповнюються відповідним середнім для праметра значенням.	Відсоток пропущених записів, що не можуть набувати значення NULL більший за Z_0
Видалення дублікатів	Числовий, строковий	*В*изначення повторюваних записів у датасеті	• Якщо дублікати за певним параметром заборонені, відбуватиметься пошук та їх видалення. Також це стосується наявність дублікатів-об'єктів загалом, або повторюваних записів • Перевірка дублікатів за первинним ключем (якщо у метаданих він наданий). Спершу перевіряються дані за зазначеними полями у ключі, а потім – усі записи без них, тобто, якщо ми вилучимо параметри ключа, чи не буде у записах повністю ідентичних даних.
Приведення текстових значень	Строковий	Текстові поля можуть приймати три класи А Б і В, для кожного рядка в датасеті це один з трьох даної колонки (А, Б, В) На виході отримуємо 3 колонки, що мають відповідні назви (А, Б, в), і для кожного рядка в датасеті це будуть нулі і одиниця в тій колонці, до якої вона належить.	Кількість текстових значень не більше певного числа N_0

Більшість систем IoT мають типові або незмінні шаблони даних. Для підвищення загального часу виконання попередньої обробки, пропонується формувати метадані вхідного датасету на етапі аналізу вхідного потоку та зберігати на сервері. Після формування алгоритму попередньої обробки на кроці «Визначення правил обробки», правила у декларативному поданні додаються до раніше сформованих метаданих та запис на сервері оновлюється.

При наступному запиті до серверу, метадані нового потоку повинні бути порівняні із вже збереженими метаданими. При знайденні співпадань, блок «Визначення правил обробки» пропускається, і очищення відбувається за вже сформованими правилами. Таким чином, при зберіганні порівняно невеликого обсягу даних, даний підхід дає значний виграш у часі обробки через підвищення рівня автоматизації.

Переваги:

1. Підвищення достовірності кінцевого результату роботи мережі IoT.
2. Підвищення якості вхідних даних за рахунок виправлення втрат та видалення хибних записів.
3. Зменшення часу на попередню обробку даних за рахунок використання шаблонів обробки.

Недоліки:

1. Використання додаткових ресурсів для проведення попередньої обробки даних.
2. Застосування додаткової пам‘яті сервера для зберігання шаблонів обробки.

Висновки. В даній роботі проаналізовано існуючі способи попередньої обробки вхідних датасетів та запропоновано зміни для підвищення ефективності та більш вигідного використання програмних та часових ресурсів. Сформовано критерії підбору алгоритмів очищення даних від випадково виникаючих помилок у заданому кроці процесу попередньої обробки 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.

Література

1. Buhaienko Y., Globa L.S., Liashenko A., Grebinichenko M. “Analysis of clustering algorithms for use in the universal data processing system”, OSTIS 2020.
2. R. Deepa. Methods and techniques to evaluate the performance of Data Cleansing Algorithms for very Large Database Systems / R. Deepa, Dr. R Manicka Chezian. // IJARCET. – 2016.
3. A Domain-Independent Data Cleaning Algorithm for Detecting Similar-Duplicates / Kazi Shah Nawaz Ripon Ashiqur Rahman and G.M. Atiqur Rahaman. // JOURNAL OF COMPUTERS. – 2010.
4. A Data Cleaning Method Based on Association Rules [Електронний ресурс] / W.Weijie, Z. Mingwei, Z. Bin, T. Xiaochun – Режим доступу до ресурсу: www.atlantis-press.com.