УДК 621.382

# АЛГОРИТМ ОЧИСТКИ ТА ОБРОБКИ ВЕЛИКИХ ДАНИХ ДЛЯ ВИРІШЕННЯ АНАЛІТИЧНИХ ЗАДАЧ

**Хрищенюк Р.А.**
*Інститут телекомунікаційних систем КПІ ім. Ігоря Сікорського, Україна*
*E-mail: khryshcheniuk@gmail.com*

**Algorithm of cleaning and processing of large data for solving analytical tasks**

The main problems of storing large data are considered. The methods of clearing and processing of large data are given.

Розглянуто основні проблеми зберігання великих даних. Наведено методи очищення та обробки великих даних.

Великі дані (Big Data) - загальна назва для структурованих і неструктурованих даних величезних обсягів, які ефективно обробляються з допомогою масштабованих програмних інструментів.

По суті поняття великих даних має на увазі роботу з інформацією величезного обсягу і різноманітного складу, вельми часто оновлюваної і знаходиться в різних джерелах з метою збільшення ефективності роботи, створення нових продуктів і підвищення конкурентоспроможності.

У великих даних є певні характеристики:

**Volume; Velocity; Variety;** Veracity; Validity; Value; Variability; Venue; Vocabulary; Vagueness.

Але три з них основні, їх ще називають «правило VVV» - три ознаки, якими великі дані повинні володіти.

**Volume** - обсяг (дані вимірюються за величиною фізичного обсягу документів).

**Velocity** - дані регулярно оновлюються, тому їх потрібно постійно оброблювати.

**Variety** - різноманітні дані можуть мати неоднорідні формати, бути неструктурованими або структурованими частково.

Програми, орієнтовані на обробку великих обсягів даних, мають справу з файлами даних об'ємом від декількох терабайт до петабайта. На практиці ці дані зберігаються в різних форматах, можуть мати пробіли, містити параметри, які можуть бути не потрібні для подальшої обробки. Обробка подібних наборів даних зазвичай відбувається в режимі поетапного аналітичного конвеєра, що включає стадії очищення, перетворення та нормалізацію даних.

По-перше, більшість анкетних даних вводяться респондентами або операторами вручну. Через неуважність або з інших причин вони допускають помилки в словах, не заповнюють обов'язкові поля анкет, скорочують назви вулиць або інших об'єктів, заносять відомості не в ті поля.

По-друге, не у всіх програмах, в які вносять відомості, налаштовані обмеження на значення, що вводяться. Наприклад, в MS Excel можна

забивати інформацію, не ставлячи навіть тип даних. Якщо ж програма розроблена фахівцями під наявне завдання, то вона включає в себе обмеження навіть на внесений формат даних або допускає введення тільки дозволених символів.

По-третє, дані не заповнені або заповнені не до кінця. Деякі поля залишаються респондентами порожніми. Також слід відмітити дублювання даних. Кілька записів мають один і той же зміст.

Тому перший і найважливіший етап, з точки зору обробки великих даних, є очищення даних.

**Очищення даних** (data cleaning, data cleansing або scrubbing) займається виявленням і видаленням помилок і невідповідностей в таблиці даних з метою поліпшення якості даних. Для очищення і спрощеної подальшої обробки даних був розроблений такий алгоритм дій (рис.1).

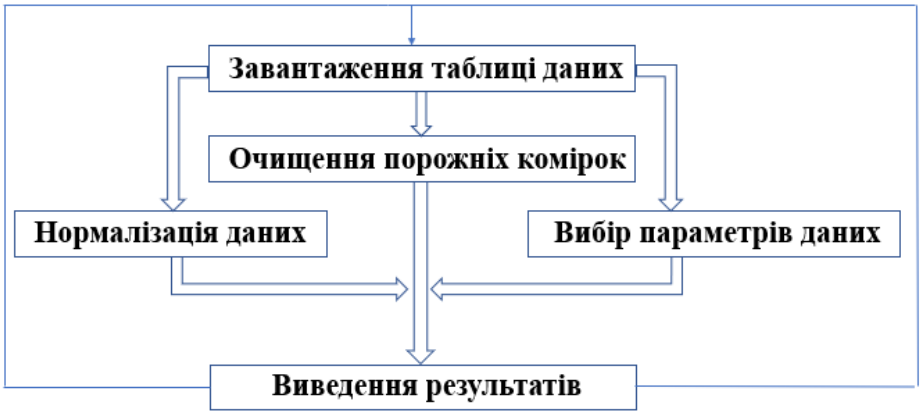


Рис. 1. Алгоритм обробки великих даних.

Спершу отримуємо таблицю даних, наприклад, в форматі xls, для цього перевіряємо формат файлу, в якому він має бути завантажений і якщо формат вірний - зберігаємо завантажену версію файлу, без змін в певній директорії, якщо ж ні то з'являється спливаюче вікно з помилкою. Потім користувач має можливість вибрати певну обробку над своїми даними.

**«Очищення порожніх комірок»** відповідає за виявлення та видалення порожніх комірок в таблиці даних. Методом перебору всіх даних в таблиці знаходимо порожні комірки і видаляємо їх. Змінену версію зберігаємо в певній директорії.

**«Вибір параметрів даних»** надає можливість користувачу обробляти саме ті параметри даних, які йому потрібно. Завантаживши таблицю даних і вибравши цей етап обробки, користувач отримує повний список параметрів даних з таблиці, після чого він має обрати ті параметри, які йому потрібні для обробки, інші параметри даних будуть видалені з таблиці. Змінену версію таблиці даних зберігаємо в певній директорії.

**«Нормалізація даних»** реалізовує нормалізацію всіх даних в певному діапазоні. Наприклад, потрібно нормалізувати всі значення в діапазоні від 0 до 1. Методом перебору всіх даних в таблиці знаходимо комірки зі

значеннями, які не відповідають потрібному діапазону. Отримуючи такі комірки, їх потрібно конвертувати в числовий формат, адже отримуємо комірку таблиці даних ми в вигляді рядка символів. Після конвертації та отримавши числове значення комірки, його потрібно нормалізувати. Для нормалізації даних найчастіше використовують метод(1)

$$\boldsymbol{y(x) = \frac{x - x_{min}}{x_{max} - x_{min}}.}$$

Отримане значення позначається «x», $x_{min}$ – мінімальне значення з таблиці даних, $x_{max}$ – максимальне значення з таблиці даних. Після пройдених дій всі дані в таблиці відповідають діапазону від 0 до 1. Змінену версію таблиці даних зберігаємо в певній директорії.

Після обробки користувач отримує результуючу таблицю даних з певними змінами і має можливість повернутись на початковий етап алгоритму, аби завантажити іншу таблицю даних. Окрім завантаження іншої таблиці даних, користувач має можливість завантажити змінені версії таблиці даних, яку він обробляв, аби повторити обробку, наприклад, з іншими параметрами даних.

**Висновок:** При розробці алгоритму ключовим етапом є розбиття загальної задачі на ряд підзадач, кожна з яких буде виконуватися окремим потоком. Потік представляє деяку частину коду програми. За допомогою багатопоточності ми можемо виділити в програмі кілька потоків, які будуть виконувати різні завдання одночасно, витрачаючи менше часу на виконання. При виконанні деяких завдань такий розподіл може досягти ефективнішого використання ресурсів комп'ютера, а також незалежність від інших підзадач. При цьому під незалежністю розуміється використання своїх власних ресурсів, мінімум в потребі синхронізації з іншими потоками. При обробці згідно даного алгоритму, користувач отримує очищену та більш структуровану таблицю даних.

Удосконалення даного алгоритму полягає в розбитті великих таблиць даних на декілька підтаблиць певного розміру, що дозволяє працювати з таблицями великих об’ємів даних та надає можливість виконувати певну обробку над потрібною частиною таблиці даних.

**Література**


1. Большие данные (Big Data) [Електронний ресурс]. – 2017. – Режим доступу до ресурсу: http://www.tadviser.ru/a/125096.
2. Чехарин Е. Е. Большие данные: большие проблемы / Е. Е, Чехарин. // Международный электронный научный журнал. – 2016.
3. Латышева А.М. Big data. Актуальность и перспективы использования / А.М. Латышева, Ю.Е. Гапанюк // ФГБОУ ВПО «МГТУ им. Н.Э.Баумана».
4. Очистка данных: проблемы и актуальные подходы [Електронний ресурс] // Журнал BPM World. – 2000. – Режим доступу до ресурсу: http://iso.ru/ru/press-center/journal/1789.phtml
5. Очистка персональных данных [Електронний ресурс] // BaseGroup Labs – Режим доступу до ресурсу: https://basegroup.ru/community/articles/person-data-part1.