НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМ. І. І. СІКОРСЬКОГО»

ННК «ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ»

Кафедра системного проектування

**Реферат**

із дисципліни «Оброблення надвеликих масивів даних»

на тему: *«Грані даних: структуровані, неструктуровані, на мовах натуральних, графові дані, аудіо- та відео дані, дані пошукові»*

**Виконала:**

студентка 5-го курсу

групи ДА-11мп

Молчанова В.С.

Київ – 2021

**Зміст**

[Вступ](#_a7sphlqd7mgo) **3**

[Структуровані дані](#_1q1hdhd6bsou) **4**

[Напівструктуровані дані](#_yagc8kerhg7g) **4**

[Неструктуровані дані](#_q2kgshju9g69) **5**

[Дані на природній мові](#_orws4s9rmbgv) **5**

[Машинні дані](#_6sbmgdhbrihn) **6**

[Графові (мережеві) дані](#_ssp9wclaza9) **6**

[Аудіо, відео та графіка](#_cgena8rae7eq) **7**

[Потокові дані](#_b1xj579ifo4q) **8**

[Процес аналізу даних](#_fr1msn9mz8ee) **8**

[Висновки](#_jl31b9cz2me0) **10**

[Джерела](#_h9k1fvlnb4qg) **12**

# Вступ

З розвитком інформаційних технологій з’явилася можливість аналізувати великі об’єми даних, що дало поштовх для стрімкого розвитку науки про дані (data science) та формування нових підходів та технології роботи з даними, що отримали узагальнюючий термін big data. Згідно оцінок IBM, в станом на 2020 рік загальний об’єм даних складає 35 зетабайт даних, а згідно передбачень експертів у 2025 році досягне 175 зетабайт [1].

Задачі роботи з великими даними, в контрасті з традиційними підходами в роботи з даними, характеризуються трьома осбливостями, які в англомовній літературі отримали назву “три V”: об’єм (volume), розмаїття (variety) та швидкість (velocity). Відповідно, для вирішення такої категорії задач змінюються підходи в зборі, зберіганні, пошуку, обміні, обробці, аналізі та візуалізації даних.

Великі дані за структурою можна поділити структуровані, неструктуровані, за форматом на мовні дані (машинні та дані на природній мові), графові, медійні (аудіо, відео та графіка, потокові та інші.

Предметом даного реферату є розгляд різних категорій даних в контекті науки про данні та big data.

# Структуровані дані

Частка структурованих даних складає 10-20% від загального об’єму даних, в той час як частка неструктурованих даних оцінюється в 80-90% [2].

Структуровані дані існують в організованому форматі та відповідають певній моделі даних, яка описує їх структуру та типи даних. Формат даних дає змогу легко отримати доступ до потрібних даних, що значною мірою спрощує аналіз.

Як правило такі дані зберігаються в базах даних (зокрема в реляційних базах даних) або в таблицях (наприклад, Google Sheets або Excel). Мова структурованих запитів (SQL) є основним інструментом для роботи з реляційними базами даних, що дає змогу отримувати дані, виконувати над ними різні маніпуляції, додавати нові дані та видаляти непотрібні.

Варто зазначити, що структуровані дані можуть містити в собі неструктуровані дані. Наприклад, форма, яка в одному з питань пропонує користувачу обрати варіант з перелічених або зазначити власний. Відповідь на таке питання буде прикладом неструктурованих даних в записі, який буде внесено до бази даних.

# Напівструктуровані дані

Велика частка даних за своєю структурою є якоюсь мірою гібридною, що породжує термін напівструктурованих даних. Такий формат включає можливість виділити ключові слова, теги та метадані в неструктурованих даних. Прикладами таких даних можуть бути текстові дані в спеціальному формат (як-то документи чи резюме), файли спеціального формату (HTML, XML, JSON), або дані, які зберігаються в NoSQL базах даних.

# Неструктуровані дані

В неструктурованих даних важко виділити якусь конкретну модель даних, адже їх зміст залежить від контексту та має варіативний характер. Зберігаються такі дані у форматі файлів у файлових системах та NoSQL базах даних.

Прикладами таких даних є зображення, відео, аудіо, книжки, електронні листи, книжки та інші.

# Дані на природній мові

Дані природною мовою є різновидом неструктурованих даних, з якими часто доводиться мати справу в області аналізу даних. Зокрема в рекомендаційних алгоритмах соціальних мереж та блогів, аналізі електронних медичних записів, аналізі коментарів на сайтах компанії для швидкого реагування, автоматичного розпізнавання важливої інформації в резюме для автоматизації роботи рекрутинга та інших сферах.

В обробці таких даних досягнуто значного успіху, зокрема в вирішенні таких завдань як: розумні асистенти (Google Assistant, Alexa або Siri), які розпізнають живе мовлення та виконують певний перелік завдань; оптимізація результатів пошуковиків (Google, Yandex), робота яких базується на тематичному аналізі сторінок для подальшого ранжування за передбаченою відповідністю теми; граматичний аналіз та машинний переклад (Grammarly, Google Translate, DeepL), задачею яких є аналіз тексту на предмет наявності граматичних помилок та граматично коректний переклад тексту; аналіз сентиментів, що допомагає класифікувати відгуки на негативні й позитивні; тематичне моделювання для групування текстів за тематикою.

Однак, моделі адаптовані для однієї предметної області можуть давати поганий результат в іншій предметній області, що вказує на дуже низьку узагальненість алгоритмів обробки природної мови.

# Машинні дані

До машинних даних відносять ту інформацію, яка автоматично генерується комп’ютером чи обчислювальною системою без втручання людини. Прикладами таких даних є логування інформації при роботі застосунків та сервісів, інформація з мережевого обладнання, записи про деталі дзвінків та інші.

В порівнянні з даними на природній мові, машині дані часто мають доволі чітку структуру, що полегшує їх аналіз. Наприклад, логування відбувається згідно спеціального патерну, зазначеного в програмі, що дає змогу автоматично пасити такі дані та зберігати їх в структурованому вигляді. Проблемою є те, що такі дані генеруються в надвеликих об’ємах, що вимагає високопродуктивних інструментів їх обробки.

# Графові (мережеві) дані

Графовими даними вважаються такі дані, які представлені у вигляді графу, мережі, за своє структурою. Відмінністю від складних реляційних баз даних є те, що головний акцент в таких даних стоїть на зв’язку між сутностями або об’єктами.

Найкращим прикладом графових даних є соціальні мережі. В ході аналізу таких даних можна обрахувати найкоротший шлях між учасниками мережі, зв’язок між двома людьми в мережі та передбачити потенційні зв’язки, що дає змогу рекомендувати учасникам соціальних мереж нових “друзів”. Такі дані можна також використовувати для рекомендації контенту, базуючись на інтересах людей, що пов’язані з користувачем.

Також широкого використання аналіз графових даних отримав в сфері кібербезпеки для виявлення ботнетів [5].

Для зберігання графових даних використовуються графові бази даних, для доступу до них спеціальні мови запитів, як-то SPARQL.

# Аудіо, відео та графіка

Аудіо, відео та зображення — доволі складний вид даних для аналізу, які є ще одним прикладом неструктурованих даних.

Аналіз аудіо даних в більшості випадків — це аналіз вербального природного мовлення, трохи рідше розпізнавання та класифікація музичних композиції (наприклад за жанрами та піджанрами, що використовується такими музичними застосунками як Spotify для рекомендацій користувачам музики), розпізнавання спеціальних сигналів (наприклад сигналізації), розпізнавання видів тварин за звуками (активно використовується в акустичній океанографії). Як правило, для аналізу природної мови використовуються рекурентні нейронні мережі, приховані марковські моделі.

Серед аналізу зображень та відео можна виділити задачі класифікації зображень (медичні візуалізації, сканери QR-кодів), виявлення об'єктів на зображеннях та відео (розпізнавання рукописних записів, розпізнавання дорожніх знаків в проектах безпілотних автомобілів), сегментація зображень, ідентифікація індивідуальних об'єктів (розпізнавання облич, або задачі біометричної аутентифікації), фіксація рухів на відеограгментах, відновлення зображень та інші. В класифікації зображень використовують згорткові нейронні мережі, для виявлення об’єктів використовують такі алгоритми як YOLO, R-CNN, Single Shot MultiBox Detector та інші.

# Потокові дані

Потокові дані можуть бути представлені у різних форматах, однак визначальною рисою є те, що вони поступають в систему при виникненні певних подій, а не завантажуються в сховища великими об’ємами. Прикладами таких даних є новини, які агрегуються та рекомендуються користувачам в розділі актуальної інформації, чи виділення популярних тем в соціальних мережах, прямі трансляції, фінансова інформація та інші.

# Процес аналізу даних

Процес аналіз даних можна описати в 6 етапів: постановка мети аналізу даних, збір даних, підготовка даних, дослідження даних, моделювання даних та відображення й автоматизація. Як правило процес аналізу даних має ітеративний характер, адже результат дослідження та моделювання, може призвести як до корекції методології підготовки даних, так і до зміни завдання й необхідності пошуку нових даних. В процесі аналізу може бути велика кількість описаних циклів.

**Формулювання мети дослідження.** На даному етапі потрібно сформувати проблему, яку потрібно вирішити, описати можливу користь дослідження, визначити необхідні ресурси та визначити методологію чи методології дослідження даних.

**Збір даних**. Дані можуть бути отримані від замовника дослідження в зібраному вигляді, або дані доведеться збирати самостійно. Можуть бути як автоматичні чи напівавтоматичні методи збору даних, так і мануальний збір даних. На даному етапі може знадобитись експертна група для розмітки даних (якщо мова йде про “навчання з учителем”).

**Підготовка даних.** На даному етапі непідготовлені дані потрібно представити в вигляді, прийнятному для обраної методології аналізу даних. Можна виділити такі етапи: очитка даних, яка полягає у видаленні некоректних та неповних даних чи корекція (наприклад, в деяких випадках частково відсутні дані можна замінити певними значеннями); інтеграція, яка полягає в об’єднанні даних з різних джерел для розширення наявної інформації; та перетворення даних, що полягає в представленні даних в форматі, придатному для подальшого аналізу.

**Дослідження даних.** На даному етапі дослідник намагається краще зрозуміти дані, дослідити залежності між змінними, оцінити розподіли даних та виявити можливі “викиди”. Для цього використовуються описові статистики та методи візуалізації.

**Побудування моделі.** На даному етапі використовуються статистичні методи та алгоритми машинного навчання для досягнення мети, поставленої перед дослідником. В процесі моделювання обираються алгоритми, підбираються гіперпараметри алгоритмів та відбувається тестування моделі.

**Відображення та автоматизація.** Результати дослідження представляються у форматі звіту (якщо задача була саме аналізу даних) або результати автоматизуються, якщо предметом дослідження було створено моделі, яку згодом потрібно буде використовувати в бізнес-процесах.

# Висновки

В даній роботі було розглянуто процес аналізу даних та проблему великих даних, як клас задач, що вимагає нових рішень, через великий об’єм даних, її гетерогенність та високу швидкість надходження нових даних.

Загалом дані можна поділити на 2 категорії: структуровані дані та неструктуровані, які за оцінками експертів складають 10-20% та 80-90% усіх існуючих даних, відповідно. Так як розподіл між цими двома категоріями не дуже чіткий, можна виділити гібридну категорію — напівструктуровані дані.

Серед текстових даних можна виділити дані на природній мові та машинні дані. Машинні дані, як правило, мають чітку структуру, одна генеруються у дуже великій кількості, що вимагає відповідних підходів до їх аналізу. Дані природною мовою — це неструктуровані дані, що робить задачу аналізу таких даних значно складнішою. Попри значні успіхи в сфері аналізу природної мови, такі моделі погано генералізуються і показують хороші результати тільки в межах предметної області, для якої вони розробляються. Однак, існуючі технології активно використовуються соціальними мережами, перекладачами, розумними асистентами та в інших прикладних сферах.

Графові дані від інших категорій відрізняє те, що велика увага приділяється саме зв’язкам між сутностями. Яскравим прикладом є соціальні мережі.

Зображення, відео та аудіо дані, як правило аналізуються за допомогою алгоритмів нейронних мереж. Попри те, що задача комп’ютерного зору є тривіальною для людини, вона є дуже складною для інформаційних систем, тому в даній сфері наразі успіхи є в вирішенні доволі вузьких задача, як-то класифікація зображень, сегментація зображень, відновлення зображень, аналіз рухів, розпізнавання та ідентифікація об’єктів.

Процес аналізу даних полягає в формулюванні задачі, зборі даних, підготовці, дослідженні даних, побудові моделі та представлення результатів. Процес має ітераційний характер.

# Джерела

1. Holly Vatter (Nov. 30, 2020). Addressing data growth with scalable, immediate and live data migration. URL: <https://www.ibm.com/blogs/journey-to-ai/2020/11/addressing-data-growth-with-scalable-immediate-and-live-data-migration/>
2. Tobias Geisler Mesevage (Aug. 26, 2020). Structured vs Unstructured Data: What’s the Difference? URL: <https://monkeylearn.com/blog/structured-data-vs-unstructured-data/>
3. Anhelina Leigh. 10 Examples of Natural Language Processing (NLP) and How to Leverage Its Capabilities. URL: <https://global.hitachi-solutions.com/blog/natural-language-processing>
4. Натан Марц, Джеймс Уоррен. Большие данные. Принципы и практика построения масштабируемых систем обработки данных в реальном времени. – Изд. дом «Вильямс», Москва, Санкт-Петербург, Киев, 2017, 368 с.
5. Ying Xing, Hui Shu, Hao Zhao, Dannong Li, Li Guo, "Survey on Botnet Detection Techniques: Classification, Methods, and Evaluation", Mathematical Problems in Engineering, vol. 2021, Article ID 6640499, 24 pages, 2021. <https://doi.org/10.1155/2021/6640499>