УДК 519.8

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ВЕЛИКИХ ДАНИХ: ПОТОЧНИЙ СТАН ТА ПЕРСПЕКТИВИ НА МАЙБУТНЄ

Онищенко А. О., аспірант, artem.onyshchenko@kname.edu.ua, ХНУМГ ім. О.М.Бекетова
Онищенко Д. О., студентка, darya.onyshchenko@kname.edu.ua, ХНУМГ ім. О.М.Бекетова
Бочаров Б. П., к. т. н., доцент, boris.bocharov@kname.edu.ua, ХНУМГ ім. О.М.Бекетова

Кількість веб-сторінок, індексованих Google, була мільйонами у 1998 році, мільярдами у 2000 році та трильйонами у 2010 році. Зі зростанням кількості користувачів Інтернету, обсяги даних також зростали. З появою різноманітних соціальних мереж, збільшення обсягів даних прискорилося з вражаючою швидкістю. Очікується, що воно буде на рівні 35% на рік та у порядку зетабайтів. Аналітика великих даних надає різноманітні уявлення про дані, які можуть допомогти науковцям та дослідникам виявити приховану інформацію в реальному часі. Дані — це кількісні показники. Аналітика даних переживає революцію з безліччю можливостей, які вона надає. Таким чином, великі дані стануть рушійною силою для бізнесів, академічної сфери, фінансових послуг та державної політики.

Сьогодні 85% генерованих даних є невпорядкованими, оскільки вони походять з різних джерел і мають різноманітні форми. Ці величезні обсяги даних важко зберігати та обробляти традиційними методами аналізу даних. Великі дані можна охарактеризувати за допомогою трьох основних характеристик:

Обсяг. Завдяки мініатюризації пристроїв Інтернету речей (IoT), щодня зафіксовано терабайти даних за допомогою датчиків. Більш ніж чотириста годин відео завантажуються на YouTube, а 320 облікових записів Twitter створюється щохвилини. Вісімсот мільйонів користувачів щодня активні на Facebook. Обсяг даних стрімко зростає.

Різноманітність. Дані генеруються з різних джерел. Вони можуть бути у вигляді твітів, листів, форм, фотографій, відео, аудіо, транзакцій тощо. Тому у даних немає конкретної структури.

Швидкість. Темп, з яким ці величезні дані генеруються, захоплюються та обробляються, є дуже високим. Фіксують поведінку користувачів зі швидкістю мільйонів подій за секунду. Повідомлення в соціальних медіа стають вірусними за лічені хвилини.

Згадані вище характеристики створюють значні труднощі для науковців у сфері даних, що намагаються виявити цінні знання з великих даних, оскільки традиційні методи аналітики даних не здатні обробляти дані з такими

особливостями. Тому процес аналізу даних потребує перегляду з точки зору обсягу, різноманітності та швидкості.

Попередня обробка даних є важливою, адже величезний обсяг даних з різними схемами важко обробляти на існуючих платформах та алгоритмах. Специфічні методи попередньої обробки необхідні для різних доменів даних, оскільки вони суттєво впливають на результати майнінгу. Задачі очищення, вибірки та стискання даних є критичними і вимагають ефективного виконання. Також висока вартість попередньої обробки для лог-даних, даних датчиків, маркетингових даних тощо, тому може бути розглянуто поділ та завоювання стратегії для попередньої обробки на різних машинах або в хмарах.

Платформа вимагає хоча б двох ресурсів: доступу до даних та обчислювальних процедур. За наявності дуже великих даних, використовуються численні вузли чи кластери для зберігання даних. Основна проблема полягає в доступі до даних з усіх вузлів під час обчислень. Викликами є безпека під час отримання даних з датчиків та їхнього обміну з іншими системами, а також різні "вузькі місця" для аналітики даних через різні вхідні системи даних.

Більшість існуючих алгоритмів дата-майнінгу призначені для централізованого обчислення і не призначені для роботи в паралельних системах. Для великих даних алгоритми можуть бути розроблені для таких категорій як: алгоритми кластеризації, алгоритми класифікації та алгоритми видобутку частих шаблонів/наборів елементів.

Виведення великих даних також є проблемою через різноманітність схем. Відображення результатів майнінгу є важливим. Були запропоновані різні бенчмарки, які охоплюють усі характеристики великих даних та різні показники продуктивності. Таким чином, потрібно звертати увагу на ці два важливих фактори при виведенні великих даних.

Окрім стандартних метрик для майнінгу даних, як-от час виконання, пропускна здатність, масштабованість та використання пам'яті, необхідно включити інші метрики. Час завантаження даних, час на запити, час процедурних запитів, залишковий час для запитів, час для мапінгу та редукування задач тощо. Затримка читання/запису та пропускна здатність важливі для платформи хмарних обчислень. Також важливі витривалість до помилок та вартість послуг у хмарах.

Візуалізація, тобто представлення результатів є значущим у аналітиці великих даних. Для якісного аналізу результатів важливо, як вони представлені користувачу. Візуалізація є складною для великих даних через великий обсяг та різноманітність розмірів даних. Деякі техніки, такі як зниження розмірності, були запропоновані, але вони не підходять для складних великих даних і можуть призвести до втрати інформації.

Проблеми великих даних відрізняються від традиційної аналітики даних, та існує багато дослідницьких викликів, які мають бути вирішені у майбутньому. Приватність є важливою, оскільки виникають питання коли системи зберігають особисту інформацію користувачів без забезпечення угоди про рівень

обслуговування. Забезпечення безпеки інформації та рівень захисту від зовнішніх інтерфейсів важливі, особливо для динамічних даних та пристроїв, які підключаються. Виклик становить і синхронізація в паралельних обчислювальних платформах. Алгоритми майнінгу даних, розроблені для процесу KDD, не є ефективними для великих даних, та потрібні нові моделі та методи для їхньої обробки, особливо для задач, де простір рішень дуже великий або проблема є NP-складною. Якість кінцевих результатів важлива, оскільки весь процес майнінгу даних має цінність лише якщо кінцевий результат генерує певну цінність. Візуалізація та злиття інформації також є важливими для якості кінцевих результатів. З точки зору платформи та рамок, акцент зміщується на продуктивність та результати, а з точки зору обчислень - великі дані та алгоритми майнінгу, такі як кластеризація, класифікація, пошук асоціативних правил, є актуальними викликами.

Висновок. Майнінг великих даних стоїть на порозі розвитку у сфері науки про дані, пропонуючи нові можливості та виклики у різноманітних інженерних областях. Незважаючи на потребу у високопродуктивних обчислювальних платформах і розробку спеціалізованих алгоритмів для кластеризації, класифікації та пошуку шаблонів, майнінг великих даних відкриває шлях до глибшого розуміння даних. Станом на зараз, він зіткнувся з проблемами приватності, безпеки та якості результатів, підкреслюючи необхідність розробки рішень у реальному часі для ефективного аналізу даних. Важливість майнінгу великих даних зростає з кожним днем, акцентуючи на потребі подальших досліджень для подолання існуючих викликів та використання майбутніх тенденцій.

Список використаних джерел

1. C. C. Aggarwal, editor. Managing and Mining Sensor Data. Advances in Database Systems. Springer, 2013.
2. Apache Hadoop, http://hadoop.apache.org.
3. P. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, and G. Lapis. IBM Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill Companies,Incorporated, 2011.