

УДК 004.6

Демченко А.С.¹, Куликовська Н.А.²

¹ студ. гр. КНТ-511м, НУ «Запорізька політехніка»

² старш. викл. НУ «Запорізька політехніка»

МЕТОДИ ОБРОБКИ BIG DATA

У сучасному світі з поширенням інтернету та нових технологій з кожною секундою генерується дедалі більше інформації. А значить, стає все більш актуальною і робота з нею – аналітика великих Big Data. Прикладами застосування аналітики великих даних є, наприклад, налаштування контекстної реклами за результатами запитів у пошукових системах, обробка даних з різних датчиків або камер спостереження та багато інших.

Вже з повсякденного досвіду відомо, що завданням аналітики великих даних є знаходження найоптимальніших і водночас найменш ресурснозалежних рішень. Виділяють чотири види аналітики Dig Data: описову, діагностичну, прогнозуючу і розпорядницьку. Залежно від мети, яку ставить перед собою аналітик, вибираються різні методи та техніки аналізу великих даних. Таких методик і техніки аналізу великих даних існує безліч.

Розглянемо основні їх:

- фільтр Блума – структура даних, яка швидко і не витрачаючи великої кількості пам'яті, визначає приналежність елемента;

- хешування – коли з даних, поданих на вхід алгоритму, виходить унікальний рядок, «хеш», завдяки якому можна відрізнити ці дані від інших. При хешуванні треба переконатися, що однакові дані, подані на вхід, видають однаковий «хеш» на виході, і прагнути того, щоб різним даним присвоювався різний «хеш», хоча цілком цю умову виконати проблематично;

- A/B тестування (іноді також називають Split URL Testing) виділяють контрольну вибірку, яку по черзі порівнюють з іншими. Її часто використовують власники сайтів із великою кількістю відвідувачів для прийняття маркетингових рішень. Шляхом застосування даної методики виявляються найбільш ефективні, оптимальні комбінації для досягнення певної мети, наприклад, просування товару або послуги;

- Data Mining, інтелектуальний аналіз даних – призначений для виявлення прихованих взаємозв'язків усередині великих баз даних. Відмінність даної технології полягає у можливості самостійно виявляти такі закономірності та вибудовувати гіпотези. До основних алгоритмів data mining відносять:

- data fusion and data integration: методики, які з'єднують і аналізують дані з різних, здебільшого, гетерогенних джерел;

- краудсорсинг – отримання даних силами великого, зазвичай, невизначеного кола осіб;

- машинне навчання - гілка штучного інтелекту, завдання якої змусити комп'ютер самостійно вчитися рішенням якоїсь конкретної мети. Такі методи аналітики поділяються наступним чином:

а) навчання з учителем (supervised learning) – наявність втручання людини в процес обробки даних;

б) навчання без вчителя (unsupervised learning) – відсутність втручання людини у процес обробки даних;

в) ensemble learning – об'єкт досліджується з кількох сторін з урахуванням статистичного аналізу чи машинного навчання отримання прогнозів по даному об'єкті.

Зрозуміло, список методів на цьому не обмежується, адже існує безліч факторів, що впливають на вибір рішення. Вибір методики, безумовно, залежить від сфери застосування. Для даних генерованих інтернетом речей властива масивна обробка поточкових даних. Для інших задач розумніше користуватися машинним навчанням чи інший методикою.

Аналітика великих даних є перспективним середовищем розвитку внаслідок великої корисності для підприємств, так суспільства в цілому. З результатами аналітики Big Data сучасна людина зустрічається щодня: починаючи від реклами товарів, які збігаються з вашими останніми запитами в інтернеті, голосових помічників, смарт-техніки так і до аналізу інформації, що надходять з датчиків протягом будь-якого експерименту.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Weikum, G. Transactional information systems: theory, algorithms, and the practice of concurrency control and recovery Текст. / G. Weikum, G. Vossen. USA: Morgan Kaufmann, 2002. - 853 с.

2. Pyle, D. Data Preparation for Data Mining: Text Текст. / D. Pyle. United States: Morgan Kaufmann, 1999 - 540 с.

3. Liu, B. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data Текст. / B. Liu. United States: Springer, 2007. - 532 с.