УДК 004.62 DOI: https://doi.org/10.30837/IYF.IIS.2024.405

ПРО ОДИН ПІДХІД ДО ОБРОБКИ ВЕЛИКИХ ОБСЯГІВ НЕСТРУКТУРОВАНИХ ДАНИХ

Полурєзов Д. С.
Науковий керівник – к.т.н., доц. Кравець Н. С.
Харківський національний університет радіоелектроніки, кафедра ПІ
м. Харків, Україна
dmytro.poluriezov@nure.ua

The issue of real-time processing of large datasets is delineated. The utilization of binary search method for handling unstructured data through parallelizing queries to remote repositories according to the number of processor cores of the mobile device is examined. This method can find its application in organizing parallel processing of substantial volumes of unstructured information on personal mobile devices operating under the iOS operating system.

Поняття Big Data з’явилось на початку XXI-го сторіччя як альтернатива традиційним системам управління базами даних. Big Data на теперішній час використовується для позначення структурованих та неструктурованих даних, що є значними за обсягом, та масштабуються горизонтально. Разом з тим слід зазначити, що великі дані – це також сукупність технологій, орієнтованих на:

– обробку дуже великих за обсягом, у порівнянні зі «стандартними» сценаріями, обсягів даних;

– опрацювання даних, що швидко надходять у дуже великих обсягах;

– операції зі структурованими та мало структурованими даними паралельно та у різних аспектах їх використання.

На теперішній час технології обробки Big Data набули поширення. Хоча проблема обробки сотень гігабайт даних вже вирішена, але дані, що були створені рік тому, набагато менш цінні, ніж дані, які було отримано протягом останньої години. Збільшення кількості потоків для паралельної обробки даних може значно підвищити продуктивність систем. Однак нескінченне збільшення кількості потоків не призведе до аналогічного збільшення продуктивності. Для оптимізації використання потоків може бути застосовано методи бінарних запитів.

Загалом будь-які дані, що можуть розглядатись як джерело великих обсягів неструктурованих даних, мають щонайменше два елементи: самі дані та їх характеристики. Формально поділимо ці об’єкти на категорії:

– сутності $\boldsymbol{e}$;

– характеристики $\boldsymbol{f}$;

– асоціації між сутностями $\boldsymbol{e}$ та характеристиками $\boldsymbol{f}$.

Наприклад:

– сутність $\boldsymbol{e}$ належить документі $\boldsymbol{f}$;

– посилання на $\boldsymbol{f}$ з’явилось у зв’язку із сутністю $\boldsymbol{e}$.

Таким чином також можна визначити:

– множину сутностей ***E*** ;

– множину характеристик ***F***;

– для кожних ***e*** та ***f*** визначено номер асоціацій між ***e*** та ***f*** як $n_{e,f}$.

Визначимо загальну кількість сутностей через $|E|$; тоді кількість характеристик можемо визначити як потужність множини $F: |F|$. Відповідно до обраних припущень можна отримати:

– для кожної характеристики ***f*** – множину $e(f) = \{e \in E: n_{e,f} > 0\}$ для усіх асоційованих з ***f*** сутностей;

– для кожної сутності ***e*** – множину $f(ef) = \{f \in EF: n_{e,f} > 0\}$ для усіх асоційованих з ***e*** характеристик.

В тому випадку, коли присутні декілька сутностей, пов'язаних з однією характеристикою об’єкту, будемо використовувати алгоритм бінарного пошуку, який дозволяє за результатами відповіді на ***q*** бінарних запитів отримати множину з $N \cdot 2^{-q}$ елементів, яка буде містити необхідний об’єкт, а кількість запитів буде обчислюватись як $q = log_2(N)$ [1].

Такий саме спосіб можна використати і для сутностей. Існує ***E*** сутностей, що містять визначену кількість інформації: $\log_2(E)$. Якщо відомо, що будь-яка сутність асоційована з певною характеристикою (існує ***e(f)*** сутностей), то кількість інформації буде визначатись як: $(|e(f)|)$.Тоді той факт, що сутність пов'язана з характеристикою ***f***, дає змогу зменшити кількість бінарних запитів до:

$$(|e(f)|) = \left(\frac{|E|}{|e(f)|}\right).$$

Кількість асоціацій також можна визначити за допомогою бінарних запитів, які необхідно сформувати для того, щоб і надалі асоціація з необхідною сутністю була відома. Кожний бінарний запит для $n_{e,f}$. зменшує кількість цих об’єктів вдвічі, а формування ***q*** запитів зменшує цю кількість до $n_{e,f} \cdot 2^{-q}$. Асоціація буде існувати до того часу, поки кількість об’єктів буде більшою за 1. Тоді найбільша кількість запитів ***q*** , для якої ще існує асоціація, можна визначити як $N \cdot 2^{-q} = 1$, що, у свою чергу, можна визначити як $q = \log_2(n_{e,f})$. Формування будь-якого додаткового запиту буде визначатись як $1 + \log_2(n_{e,f})$.

На підставі викладеного характеристики ***f*** для сутності ***e*** можна визначати як:

$$\left(\frac{|E|}{|e(f)|}\right)$$

з фактором важливості $1 + \log_2(n_{e,f})$. Враховуючи це, результуючу кількість інформації можна представити виразом наступного виду:

$$(n_{e,f})) * \left(\frac{|E|}{|e(f)|}\right) \quad (1)$$

Формула (1) є одним з варіантів визначення для кожної сутності ***e*** важливості ***I (e, f)*** у відповідності до різних характеристик ***f*** .

Виконаємо нормалізацію значення важливості:

$$V(e,f) = \frac{(1+(n_{e,f}))*\left(\frac{|E|}{|e(f)|}\right)}{\sqrt{\sum \left((1+(n_{e,f}))*\left(\frac{|E|}{|e(f)|}\right)\right)^2}}. \quad (2)$$

Кожна із сутностей ***e*** має вагу ***V (e, f)***, тому мірою наближення об'єкту E_i до об'єкту E_n слід вважати відстань між відповідними векторами $V(e, f_i)$ та $V(e, f_n)$.

Для кожної ваги ***V (e, f)***, що репрезентує визначену кількість бітів, відстань між сутностями e_1 та e_2 буде обчислюватись у відповідності до:

$$d(e_1, e_2) = \sum_{f \epsilon F} \left|V(e_1, f) - V(e, f_2)\right|. \quad (3)$$

Ця відстань залежить від кількості характеристик: якщо, наприклад, крім самих документів ми зберігаємо ще їх копії, то відстань збільшується вдвічі. Виконаємо нормалізацію відстані $d(e_1,e_2)$ в інтервалі [0,1] шляхом її ділення на максимально можливе значення цієї відстані.

В умовах невизначеності, коли значення ***A*** та ***B*** невідомі, а є лише верхні межі цих величин $\underline{a}$ та $\underline{b}$, то найбільша можлива різниця буде становити $max(\underline{a}, \underline{b})$, а саме [2]:

– якщо $\underline{a} \leq \underline{b}$, то $|\underline{a} - \underline{b}| = \underline{b} - \underline{a} \leq \underline{b}$ та $|\underline{a} - \underline{b}| \leq$ **max** ($\underline{a}$, $\underline{b}$);

– якщо $\underline{b} \leq \underline{a}$, то $|\underline{a} - \underline{b}| = \underline{a} - \underline{b} \leq \underline{a}$ та $|\underline{a} - \underline{b}| \leq$ **max** ($\underline{a}$, $\underline{b}$),

тобто в обох випадках виконується $|\underline{a} - \underline{b}| \leq$ **max** ($\underline{a}$, $\underline{b}$).

Межа $max\ (\underline{a}, \underline{b})$ досягається у двох випадках:

– якщо $\underline{a} \leq \underline{b}$, то при ***a*** = 0, ***b*** = $\underline{b}$;

– якщо $\underline{b} \leq \underline{a}$, то при ***a*** = $\underline{a}$, ***b*** = 0.

Наведена модель буде використана для побудови та подальшого дослідження методів паралельної обробки великих обсягів неструктурованої інформації шляхом розпаралелювання та оптимізації кількості запитів до сховищ за допомогою персональних мобільних пристроїв під управлінням операційної системи iOS.

Список використаних джерел:

1. A. Khovrat, V. Kobziev, A. Nazarov and S. Yakovlev. Parallelization of the VAR Algorithm Family to Increase the Efficiency of Forecasting Market Indicators During Social Disaster. / Information Technology and Implementation (IT&I-2022), November 30 – December 02, 2022, Kyiv, Ukraine. – 12 pp. CEUR Workshop Proceedings, 2022, 3347, pp. 222–233.

2. N. Shakhovska, S. Fedushko, M. Greguš, N. Melnykova, I. Shvorob and Y. Syerov: Big data analysis in development of personalized medical system. Procedia Comput Sci. 160:229–234. 2019.