DOI: 10.34185/1991-7848.itmm.2020.01.026

МАТЕМАТИЧНІ МЕТОДИ СКОРОЧЕННЯ ПРОСТОРУ АНАЛІЗОВАНИХ СТАНІВ ПРИ ОБРОБЦІ «ВЕЛИКИХ ДАНИХ»

Сироткіна О.І., к.т.н., доцент, Алексєєв М.О., д.т.н., професор,
Удовик І.М., к.т.н., доцент

НТУ «Дніпровська політехніка», Україна

Ключові слова: ВЕЛИКІ ДАНІ, СТРУКТУРА ОРГАНІЗАЦІЇ ДАНИХ, ВПОРЯДКОВАНА МНОЖИНА ДОВІЛЬНОЇ ПОТУЖНОСТІ, *M*-АРНІ КОРТЕЖІ, МЕТОД СКОРОЧЕННЯ ПРОСТОРУ АНАЛІЗОВАНИХ СТАНІВ.

Вступ. На сучасному етапі створення та застосування інформаційно-технологічних рішень у різних галузях промисловості та енергетики актуальними є завдання обробки, зберігання, аналізу та управління великими даними в умовах тимчасових, обчислювальних та інформаційних обмежень [1–8]. Однією з найважливіших проблем застосування методів роботи з великими даними є їх обчислювальна осяжність [1–3, 6].

Якщо зі збільшенням кількості оброблюваних даних, число операцій зростає експоненційно, то таке явище називається «комбінаторним вибухом» [1]. Одним з напрямків вирішення даної проблеми є використання методів скорочення простору аналізованих станів при роботі з великими даними [2, 8, 9].

Основний матеріал. Метою дослідження в даній роботі є мінімізація тимчасових і обчислювальних ресурсів при роботі з великими даними шляхом розробки математичних методів скорочення простору аналізованих станів для структури організації даних (СОД) типу «*m*-арні кортежі на основі впорядкованих множин довільної потужності (ВМДП)». Дана впорядкована структура в загальному вигляді описує шаблонний булеан – впорядковану множину всіх підмножин впорядкованої базової множини довільної потужності для будь-якого типу даних. Скорочення простору аналізованих станів виконується з використанням аналітичних функцій між елементами СОД, які були виведені на основі аналізу властивостей СОД типу «*m*-арні

кортежі на основі ВМДП» у залежності від місця розташування елементів СОД в упорядкованому булеані.

Більш детальний опис основних термінів та визначень, властивостей і математичних методів роботи з СОД типу «*m*-арні кортежі на основі ВМДП» наведено в роботах [10 – 12].

Метод скорочення простору аналізованих станів з використанням аналітичних функцій між елементами СОД базується на основній властивості СОД: кожен унікальний *m*-арний кортеж $y^n_{m,j}$ у складі булеана 2^X, який формується з елементів упорядкованої базової множини *X* потужності *n*, може бути однозначно визначеним парою індексів (*j*, *m*), де

$y^n_{m,j}$ – *m*-арний кортеж, елемент булеана 2^X;

n – потужність впорядкованої базової множини *X*;

m – довжина кортежу;

j – індекс (порядковий номер) *m*-арного кортежу у впорядкованій множині Y^n_m .

Основна ідея пропонуємого підходу полягає в необхідності розробки математичних методів рішення системи (1) для будь-якої коректної комбінації значень параметрів: n, m_1, m_2, j_1, j_2 з мінімізацією тимчасових і обчислювальних ресурсів на обробку та аналіз даних.

$$\begin{cases} y^n_{m,j} = (y^n_{m_1,j_1} \quad op \quad y^n_{m_2,j_2}) \neq \varnothing, \\ op \in \{\subset, \cap, \cup, \setminus\}, \\ 1 \le m_1 \le m_2 \le n, \\ 1 \le j_1 \le \begin{pmatrix} n \\ m_1 \end{pmatrix}, \\ 1 \le j_2 \le \begin{pmatrix} n \\ m_2 \end{pmatrix} \end{cases} \tag{1}$$

Розв'язання системи (1) здійснюється шляхом виведення набору функціональних залежностей (2) на основі аналізу властивостей СОД при різних початкових умовах, тобто для будь-яких можливих коректних комбінацій значень аргументів f_γ

$$F = \{f_\gamma(n, m_1, m_2, j_1, \eta)\} \tag{2}$$

де розташування елемента $y^n_{m_1,j_1}$ у кортежі $y^n_{m_2,j_2}$.

У роботах [10 - 12] були розглянуті математичні методи виведення набору функціональних залежностей і визначення істинності системи (1) при заданих початкових умовах: m_1=1; j_1= {1, 2, 3}.

Для операції включення визначимо частку комбінацій операндів, що представлені елементами булеана з довжинами кортежів [m_1; $m_2 > m_1$], для яких виконується умова $(y^n_{m_1,j_1} \subset y^n_{m_2,j_2}) = true$, відносно до загальної кількості комбінацій операндів з довжинами кортежів [m_1; $m_2 > m_1$] (див. формулу 3).

$$\Delta^n_{m_1,m_2} = \frac{\binom{n}{m_1} * \binom{n-m_1}{m_2-m_1}}{\binom{n}{m_1} * \binom{n}{m_2}} * 100\% \qquad (3)$$

У таблиці 1 наведено результати розрахунку $\Delta^7_{m_1,m_2}$.

Таблиця 1 – Результати розрахунку $\Delta^7_{m_1,m_2}$

m_1	m_2	Δ, %	m_1	m_2	Δ, %
1	*2*	*3*	*1*	*2*	*3*
1	1	25	3	4	11,4
	2	28,6		5	28,6
	3	42,9		6	57,1
	4	57,1		7	100
	5	71,4	4	4	5,6
	6	85,7		5	14,3
	7	100		6	42,9
2	2	9,1		7	100
	3	14,3	5	5	9,1
	4	28,6		6	28,6
	5	47,6		7	100
	6	71,4	6	6	25
	7	100		7	100
3	3	5,6	7	7	100

Графік $\Delta^7 = f(m_1, m_2)$ представлено на рис. 1.

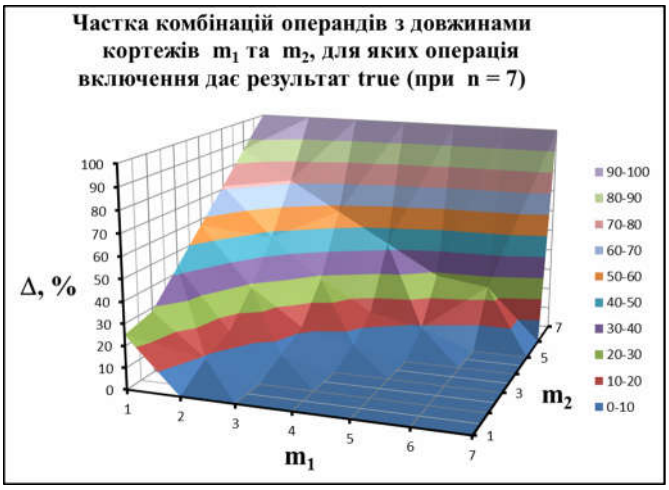

Рисунок 1 – Графічне представлення $\Delta^n_{m_1,m_2}$ при n=7

Як показано на графіку, частка комбінацій операндів, де один з них є підмножиною іншого, збільшується при збільшенні різниці $m_2 - m_1$.

Висновки. Розглянуті в статті методи роботи з СОД типу «m-арні кортежі на основі ВМДП» дозволяють отримувати результати операцій над елементами СОД з використанням набору виведених функцій (див. формулу 2) в залежності від місця розташування операндів у структурі [10–12].

У порівнянні з відомими [1-9] методами та алгоритмами роботи з великими даними, оцінка часу отримання результату змінюється з кубічної $O(n^3)$ на лінійну $O(n)$ [10–12]. Даний підхід дозволяє мінімізувати тимчасові та обчислювальні ресурси, що задіяні при обробці даної СОД, до масштабу реального часу.

References

1. A. Gaur, “Search techniques to contain combinatorial explosion in artificial intelligence,” International Journal of Engineering Research & Technology, vol. 1, issue 7, pp. 1–7, September 2012.

2. S. Yadav, A. Phulre, M. Pradesh, “A literature review on Big Data reduction methods,” International Journal of Electrical, Electronics and Computer Engineering, pp. 92–99, June 2017.

3. H. Hashem, D. Ranc, “An integrative modeling of Big Data processing,” International Journal of Computer Science and Applications, ©Technomathematics Research Foundation, vol. 12, pp. 1–15, January 2015.

4. K. Tadist, S. Najah, N. Nikolov, F. Mrabti, A. Zahi, “Feature selection methods and genomic Big Data: a systematic review,” Journal of Big Data, pp. 1–24, August 2019.
5. N. Shakhovska, O. Veres, M. Hirnyak, “Generalized formal model of Big Data,” Econtechmod. An International Quarterly Journal, vol. 5, pp. 33–38, February 2016.
6. B. Suvarnamukhi, M. Seshashayee, “Big Data concepts and techniques in data processing,” International Journal of Computer Sciences and Engineering, vol. 6, Issue-10, pp. 712–714, Oct 2018.
7. Y. Ishizuka, W. Chen, I. Paik, “Workflow transformation for real-time Big Data processing,” IEEE International Congress on Big Data, pp. 31–318, 2016.
8. I. Bifulco, S.Cirillo, “Discovery multiple data structures in Big Data through global optimization and clustering Methods,” IEEE 22nd International Conference Information Visualization, pp. 117–121, 2018.
9. K. Tasdemir, E. Merenyi, “Exploiting data topology in visualization and clustering of self-organizing maps,” IEEE Transactions on Neural Networks, vol. 20, pp. 549–562, April 2009.
10. O. Syrotkina, M. Alekseyev, V. Asotskyi, and I. Udovyk, “Analysis of how the properties of structured data can influence the way these data are processed,” Naukovyi Visnyk NHU, Dnipro, vol. 3 (171), 2019, pp. 119–129.
11. Syrotkina O. Graphical and Analytical Methods for Processing “Big Data” Based on the Analysis of Their Properties Model / O. Syrotkina, M. Alekseyev, I. Udovyk // Системні технології. Регіональний міжвузівський збірник наукових праць. – Випуск 3 (122). – Дніпро, 2019. – С. 78-90.
12. O. Syrotkina, M. Alekseyev, L. Meshcheriakov, and B. Moroz, “Methods of working with “big data” based on the application of “*m*-tuple” theory,.” Computer-Integrated Technologies: Education, Science, Production, Lutsk, vol. 36, 2019, pp. 140–152.

MATHEMATICAL METHODS FOR REDUCING THE SPACE OF ANALYZED STATES WHEN PROCESSING BIG DATA

Olena Syrotkina, Mykhailo Aleksieiev, Iryna Udovyk

Abstract. This paper addresses the problem of creating mathematical methods to optimize time and computing resources when processing Big Data. These methods are based on the proposed data organizational structure called “*m*-tuples based on ordered sets of arbitrary cardinality”. We formulated certain properties of the given data organizational structure as a consequence of the logical rules applied for the formation of *m*-tuples. A set of functional dependencies was also derived between

m-tuples based on their location in the structure. A graphical interpretation was presented to illustrate the change of dynamics in fractions of operand combinations for which one tuple is a subset of the other. It takes into account the variation in the lengths of operand tuples. We also obtained logical conclusions about the influence of the properties studied and mathematical methods of working with the given structure to minimize the computing resources involved.

Keywords: BIG DATA, DATA ORGANIZATIONAL STRUCTURE, ORDERED SET OF ARBITRARY CARDINALITY, *M*-TUPLES, METHOD FOR REDUCING THE SPACE OF ANALYZED STATES.