УДК 004.62 DOI:10.31673/2412-4338.2021.021523

Гордійчук-Бублівська О.В., Бешлей М.І., Кирик М.І., Климаш М.М. *Національний університет "Львівська політехніка", Львів*

ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ ОБРОБЛЕННЯ ВЕЛИКИХ ОБСЯГІВ ІНФОРМАЦІЇ З ВИКОРИСТАННЯМ МЕТОДУ РОЗПОДІЛЕНОГО АНАЛІЗУ ДАНИХ

Анотація: *Для створення ефективних систем оброблення даних слід використовувати різноманітні методи збору, зберігання та аналізу інформації. Щоб вирішити проблеми пошуку потрібної інформації в великих масивах даних, застосовують алгоритми машинного навчання. Враховуючи, що більшість сучасних масштабних інформаційних систем використовують велику кількість обчислювальних пристроїв, значно ефективніше застосовувати розподілені технології оброблення даних. Зокрема, широко використовується розподілене машинне навчання, в якому пристрої тренуються на локальних наборах даних і надсилають глобальній моделі тільки результати роботи. Такий підхід дозволяє підвищити надійність і конфіденційність даних, оскільки інформація про користувача залишається на одному пристрої. У статті також наведено підхід для аналізу великих обсягів інформації за допомогою алгоритму сингулярної декомпозиції даних. Даний алгоритм дозволяє як зменшити обсяг інформації, відкинувши надлишковість, так і передбачати події на основі виявлених закономірностей в даних. Визначено основні особливості розподіленого аналізу даних, можливості використання складних алгоритмів аналізу інформації та машинного навчання в таких системах. Проте, алгоритм сингулярного аналізу даних досить складно реалізувати з врахуванням розподіленої архітектури. Для підвищення ефективності використання даного методу в розподілених системах пропонується спеціальний модифікований алгоритм FedSVD. На основі цього алгоритму дані користувачів збираються з різних пристроїв, проте додається можливість додатково захистити їх від можливого втручання чи перехоплення. Результати роботи можуть бути використані при проектуванні систем для аналізу даних, збільшення надійності використання користувацької інформації, в тому числі в корпоративних інформаційних системах, фінансовій чи IT сферах тощо. Запропоновані підходи можуть слугувати основою для розробки інформаційних технологій автоматичного надання користувачам рекомендацій, передбачення аварійних ситуацій на підприємствах.*

Ключові слова: *великі дані, розподілені системи, машинне навчання, підвищення надійності оброблення інформації.*

Hordiichuk-Bublivska O.V., Beshley M.I., Kyryk M.I., Klymash M.M. *Lviv Polytechnic National University, Lviv*

IMPROVING THE EFFICIENCY OF PROCESSING BIG DATA USING THE DISTRIBUTED DATA ANALYSIS METHOD

Abstract: *Various methods of collecting, storing, and analyzing information should be used to create efficient data processing systems. To solve the problem of finding the necessary information in large data sets, machine learning algorithms are used. Given that most modern large-scale information systems use a huge number of computing devices, it is much more efficient to use distributed data processing technologies. In particular, distributed machine learning is widely used, in which devices are trained on local datasets and send only results to the global model. This approach improves the reliability and confidentiality of data because user information remains on the same device. The article also presents an approach for the analysis of large amounts of information using the algorithm of Singular Value Decomposition (SVD). This algorithm allows both to reduce the amount of information, discarding redundancy, and to predict events based on the identified patterns in the data. The main features of distributed data analysis, the possibility of using complex algorithms for information analysis, and machine learning in such systems are identified. However, the algorithm of Singular Value Decomposition is quite difficult to implement given the distributed architecture. To improve the efficiency of this method in distributed systems, a special modified FedSVD algorithm is*

proposed. Based on this algorithm, user data is collected from different devices, but the ability to further protect them from possible interference or interception is added. The results of the work can be used in the design of systems for data analysis, increasing the reliability of the user information used, including in corporate information systems, financial or IT areas, etc. The proposed approaches can serve as a basis for the development of information technology for automatic provision of recommendations to users, prediction of emergencies in enterprises.

***Keywords**: Big Data, distributed systems, machine learning, increasing the reliability of information processing.*

Гордийчук - Бубливска А.В., Бешлей М.И., Кирик М.И., Климаш М.М. *Национальный университет "Львовская политехника", Львов*

ПОВЫШЕНИЕ ЭФФЕКТИВНОСТИ ОБРАБОТКИ БОЛЬШИХ ОБЪЕМОВ ИНФОРМАЦИИ С ИСПОЛЬЗОВАНИЕМ МЕТОДА РАСПРЕДЕЛЕННОГО АНАЛИЗА ДАННЫХ

***Аннотация:** Для создания эффективных систем обработки данных следует использовать разнообразные методы сбора, хранения и анализа информации. Чтобы решить проблемы поиска нужной информации в больших массивах данных, применяют алгоритмы машинного обучения. Учитывая, что большинство современных масштабных информационных систем используют большое количество вычислительных устройств, значительно эффективнее применять распределенные технологии обработки данных. В частности, широко используется распределенное машинное обучение, в котором устройства тренируются на локальных наборах данных и отправляют глобальной модели только результаты работы. Такой подход позволяет повысить надежность и конфиденциальность данных, поскольку информация о пользователе остается на одном устройстве. В статье также приведен подход для анализа больших объемов информации с помощью алгоритма сингулярной декомпозиции данных. Данный алгоритм позволяет как уменьшить объем информации, отбросив избыточность, так и предвидеть события на основе найденных закономерностей в данных. Определены основные особенности распределенного анализа данных, возможности использования сложных алгоритмов анализа информации и машинного обучения в таких системах. Однако, алгоритм сингулярного анализа данных достаточно сложно реализовать с учетом распределенной архитектуры. Для повышения эффективности использования данного метода в распределенных системах предлагается специальный модифицированный алгоритм FedSVD. На основе этого алгоритма данные пользователей собираются с различных устройств, однако добавляется возможность дополнительно защитить их от возможного вмешательства или перехвата. Результаты работы могут быть использованы при проектировании систем для анализа данных, увеличение надежности использования пользовательской информации, в том числе в корпоративных информационных системах, финансовой или ИТ сферах. Предложенные подходы могут служить основой для разработки информационных технологий автоматического предоставления пользователям рекомендаций, предсказания аварийных ситуаций на предприятиях.*

***Ключевые слова:** большие данные, распределенные системы, машинное обучение, повышение надежности обработки информации.*

Вступ

Постійне зростання обсягів інформації, що потребує оброблення вимагає впровадження нових методів її аналізу. Оскільки за останні два роки багато сфер життя почали активно впроваджувати онлайн-послуги, проблема швидкого опрацювання даних є надзвичайно важливою. В медицині, освіті, фінансовій та економічній сферах слід забезпечити надійність користувацької інформації, а також якісну і швидку обробку [1-2].

Для покращення ефективності оброблення великих даних застосовують спеціальні алгоритми її аналізу. З їх допомогою можна визначити неважливі та повторювані дані, які потім відкидають. Також часто для корпоративних інформаційних систем доцільно проводити аналіз даних про своїх клієнтів, про популярність певних товарів чи послуг, для покращення маркетингу та підвищення прибутків [3].

Для підвищення точності алгоритмів оброблення даних їх модифікують відповідно до поставлених завдань. Зокрема, для одних систем найважливішим показником якості роботи є швидка обробка даних, для інших – безпека та конфіденційність користувацької інформації, захищеність транзакцій [4]. В роботі досліджено модифікований метод сингулярної декомпозиції даних, що дозволяє проводити аналіз великих обсягів інформації, проводити передбачення щодо майбутніх даних та при цьому підвищити конфіденційність і надійність.

Основна частина

Метою даної роботи є розробка підходів для оброблення великих обсягів інформації в комерційних інформаційних системах, швидкого аналізу та структуризації даних. Для досягнення поставленої мети проведено аналіз існуючих методів і підходів до опрацювання великих даних. Визначено основні проблеми, з якими стикаються при використанні розподілених і нерозподілених технологій оброблення інформації. В даній роботі розроблено модель оброблення даних різних обсягів на основі використання розподіленого та нерозподіленого алгоритмів аналізу даних.

Дані з корпоративних інформаційних систем постійно збираються та аналізуються. Це відбувається для покращення якості обслуговування клієнтів. До прикладу, коли користувач інтернет-магазину заходить вперше на сайт, йому пропонують зареєструватися. Після внесення даних в системі аналізуються особливості даного користувача. В залежності від його соціального статусу, віку, країни проживання пропонуються найбільш імовірно цікаві товари [5].

Завдання підбору подібних товарів чи послуг вирішують спеціальні рекомендаційні системи. В них всі дані про користувачів та товари збираються в таблиці рекомендацій. В таких таблицях розміщені відомості про клієнтів корпоративних інформаційних систем, товари чи послуги. Проте найважливішим елементом таблиць рекомендацій є оцінка товару чи послуги від користувача. Для визначення середньої оцінки аналізуються різні фактори:

- придбання товару;
- пошук на сайті;
- реакція на допис про товар в соціальних мережах (вподобання, збереження, коментарі).

Рекомендаційні системи різних компаній можуть обмінюватися даними про популярність товарів чи послуг, закономірності поведінки різних типів користувачів. Тому важливим завданням є забезпечити конфіденційність персональних даних в процесі підвищення ефективності надання таких рекомендацій.

Існує два основні типи рекомендаційних систем:

1) Орієнтовані на товар. В таких система аналізується взаємозв’язок між різними товарами. Якщо два товари подібні за певними ознаками, то доцільно їх згрупувати. Потім у випадку, якщо покупець купив один товар з групи, досить імовірно, що його зацікавлять подібні.

2) Орієнтовані на клієнтів. В таких рекомендаційних системах групують користувачів за певними ознаками. До прикладу, групі студентів туристична фірма може надсилати рекламні пропозиції про недорогі подорожі, особливо під час канікул. Людям з високим фінансовим становищем натомість пропонують престижні відпочинкові комплекси, високий рівень сервісу.

Для проведення аналізу таких таблиць рекомендацій слід використовувати математичні методи оброблення інформації. Якщо в невеликих корпоративних системах таблиці даних є невеликими, і їх можна обробити навіть без використання потужних обчислювальних пристроїв та вдосконалених алгоритмів, то для великомасштабних систем потрібно впроваджувати більш ефективні методи.

Одним з найпростіших у реалізації є алгоритм сингулярного розкладу даних (Singular Value Decomposition, SVD). Даний алгоритм збирає дані про користувачів, товари та взаємозв’язок між ними в таблиці рекомендацій. Потім дані аналізуються, визначається

залежність між групами товарів чи послуг. На основі цього надаються рекомендації про нові послуги.

Ще однією важливою проблемою оброблення великих таблиць рекомендацій є те, що часто вони досить розріджені, тобто заповнені не всі клітинки. Також таблиці рекомендацій містять дані, що мало впливають на процес аналізу. Це повторювана інформація, неважлива тощо.

Робота алгоритму сингулярної декомпозиції наведена на рис.1.

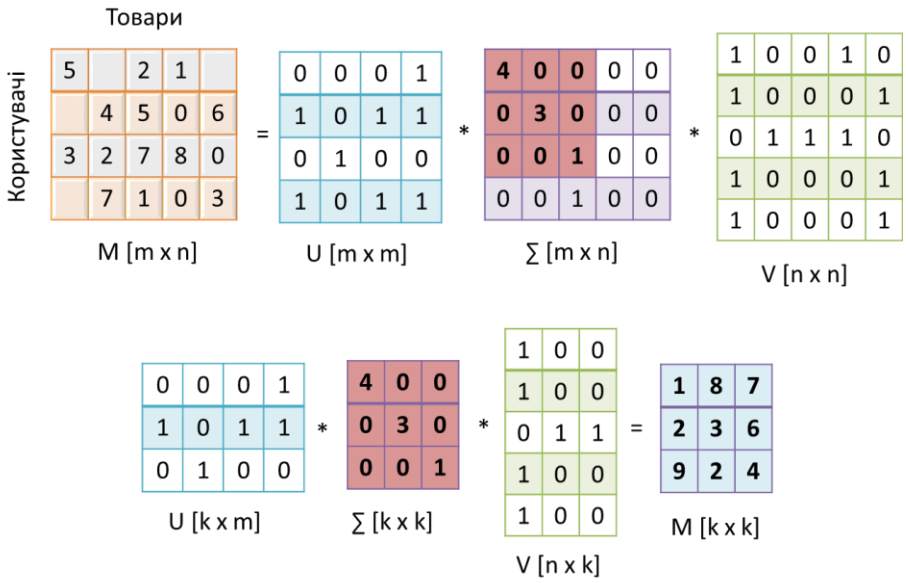

Рис.1. Алгоритм сингулярного розкладу даних

Як бачимо з рисунку 1, вхідна матриця даних може бути представлена як:

$$A = U \cdot \sum \cdot V^T \ , \qquad (1)$$

де: матриця $U(m\ x\ m)$ містить характеристики користувачів системи;

матриця $V(n\ x\ n)$ містить характеристики товарів;

матриця $\sum(m\ x\ n)$ містить залежності між користувачами і товарами.

Якщо необхідно зменшити розмір матриці M, слід провести її сингулярний розклад. Матриці U, V є ортогональними між собою. Матриця $\sum$ є одиничною, тобто всі елементи, крім розміщених на діагоналі, дорівнюють нулю. На цій діагоналі видно, що деякі початкові елементи мають додатне числове значення, а значення наступних прямують до нуля.

Можна виокремити k найбільших діагональних елементів матриці $\sum$. Потім, якщо ми оберемо в матрицях U, V і $\sum$ тільки перші k рядків та перемножимо їх, то результуюча матриця A_1 буде повторювати основні характеристики матриці A, проте без надлишкової інформації вона матиме менші розміри [7].

За допомогою алгоритму сингулярного розкладу можна значно спростити обчислення рекомендацій в великих корпоративних інформаційних системах. Тому його використовують для зменшення обсягів даних, пошуку закономірностей між ними тощо. Зокрема, даний алгоритм може використовуватися в технології Інтернету Речей. Оскільки потрібно збирати дуже багато даних з різних пристроїв та датчиків, доцільно їх оптимізувати. Також сингулярний аналіз дозволяє виявляти можливі збої в роботі обладнання, передбачати аварійні ситуації внаслідок постійного пошуку взаємозв'язків між даними, що надходять з системи.

Розподілені системи оброблення даних. Для збільшення ефективності обчислень застосовують розподілені системи оброблення даних, в яких всі задачі виконуються паралельно на декількох обчислювальних пристроях. В таких системах деякі пристрої є керуючими, вони визначають завдання, розподіляють їх на виконання між іншими пристроями, які є виконавцями [8-10].

Розподілені системи використовуються в багатьох галузях. Зокрема, модифікувалося і машинне навчання. В розподіленому машинному навчанні беруть участь багато пристроїв з власними наборами тренувальних даних. Пристрої навчаються на власних наборах даних і надсилають глобальній моделі тільки результати. Потім глобальна модель навчання оновлюється та надсилається всім іншим пристроям в мережі. Подібним чином працюють різноманітні програми чи додатки, які надають персоналізовані рекомендації щодо харчування, режиму тренувань тощо. Подібний підхід дозволяє користуватися всіма перевагами машинного навчання, при цьому не пересилати приватні дані, що підвищує їх конфіденційність та зменшує навантаження на канали передавання інформації.

Архітектуру системи з розподіленим машинним навчанням зображено на рис.2.

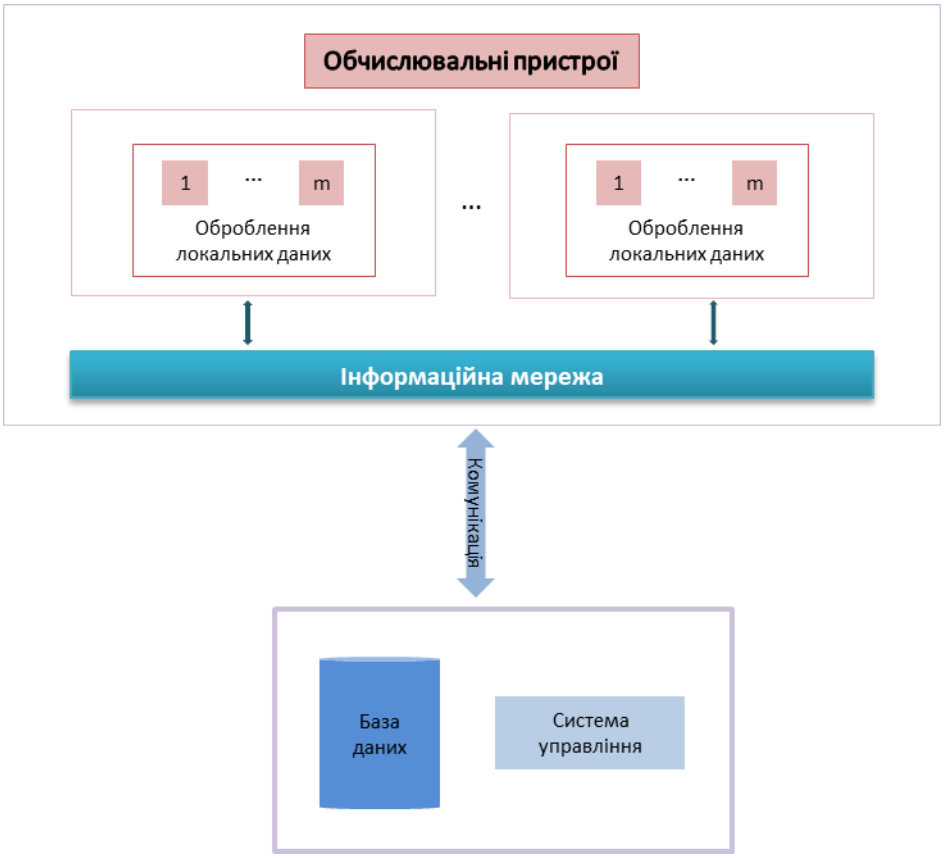

Рис.2. Архітектура системи з розподіленим машинним навчанням

Як зображено на рисунку 2, подібна розподілена організація обчислень вимагає додаткового пристосування існуючих алгоритмів. Зокрема, метод SVD, через власну складність обчислень, досить важко привести до розподіленого вигляду, оскільки всі операції виконуються послідовно та залежать одна від одної. Хоча для обчислень великих даних в великомасштабних інформаційних системах дуже важливим є застосування саме сингулярного методу, виникає проблема при зборі даних для обчислень зі значної кількості пристроїв, оскільки слід забезпечити їх надійність та конфіденційність.

Тому, пропонується модифікований алгоритм SVD – FedSVD (Federated SVD, розподілений SVD). В ньому обчислення проводяться на центральному пристрої за допомогою стандартного методу SVD. На кожному пристрої перед відправкою персональних даних відбувається генерування спеціальної маски – даних, що обчислюються на основі інформації користувача та випадкового числа, відомого тільки йому. Потім ця маска додається до даних так, щоб їх змінити. Далі блоки зашифрованих даних надсилаються глобальній моделі, яка збирає їх в одну велику матрицю та обчислює сингулярний розклад. Результатом обчислень є матриця M_1, яка поділяється на менші підматриці, що потім розсилаються назад користувачам.

Кожен пристрій отримує підматрицю m_i, яка є добутком матриць U, ∑ (вони є відомими) та V_i, яка є захищена маскою. Щоб отримати її правильне значення, до V_i застосовується зворотна операція додавання маски. Після перемножування трьох підматриць отримується результат виконання сингулярного розкладу для конкретного набору даних (рис.3).

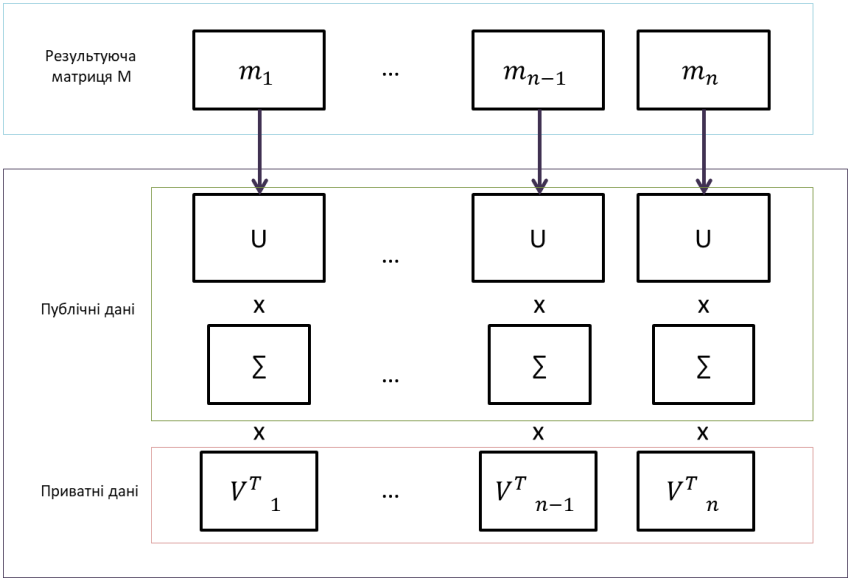

Рис.3. Алгоритм розподіленого сингулярного розкладу

Математичне пояснення роботи алгоритму FedSVD. Оскільки сингулярний розклад має вигляд:

$$M = U * \sum * V^T, \tag{2}$$

де

$$M \in R^{m\,x\,n},$$
$$U \in R^{m\,x\,m}, \sum \in R^{m\,x\,m}, V^T \in R^{m\,x\,n}. \tag{3}$$

Для розподіленої обробки розділимо вхідну матрицю на частини:

$$M = \{m_1, m_2, \dots, m_k\}. \tag{4}$$

Отже, вхідна матриця має розмірність $M \in R^{m\,x\,n}$, причому:

$$n \in \sum_i n_i, \tag{5}$$

де

$$1 \le i \le k. \tag{6}$$

Можна представити дані у вигляді:

$$M = \{m_1, m_2, \dots, m_i\}, \tag{7}$$
$$m_i = U * \sum * V_i^T. \tag{8}$$

Похідні матриці мають розмірність:

$$U \in R^{m\,x\,m}, \tag{9}$$
$$\sum \in R^{m\,x\,m}, \tag{10}$$
$$V_i^T \in R^{m\,x\,n_i}. \tag{11}$$

Фінальна матриця, яка використовується в алгоритмі FedSVD:

$$M = U * \sum * V_{\sum i}^{T}. \quad (12)$$

Порівняння ефективності роботи розподіленого та нерозподіленого алгоритму сингулярної декомпозиції даних. Для того, щоб порівняти переваги і недоліки розподіленого і нерозподіленого алгоритмів SVD, було створено програмну модель на мові програмування Python, яка дозволяє отримати графіки ефективності роботи.

Для експериментального дослідження було використано набори даних про користувачів системи рекомендацій фільмів різного обсягу. Щоб проводити обчислення використано вбудовану бібліотеку Python – TruncatedSVD, яка дозволяє проводити сингулярний розклад даних.

Залежність тривалості обчислень від обсягу обчислюваних даних для розподіленого і нерозподіленого алгоритмів SVD наведено на рис.4.

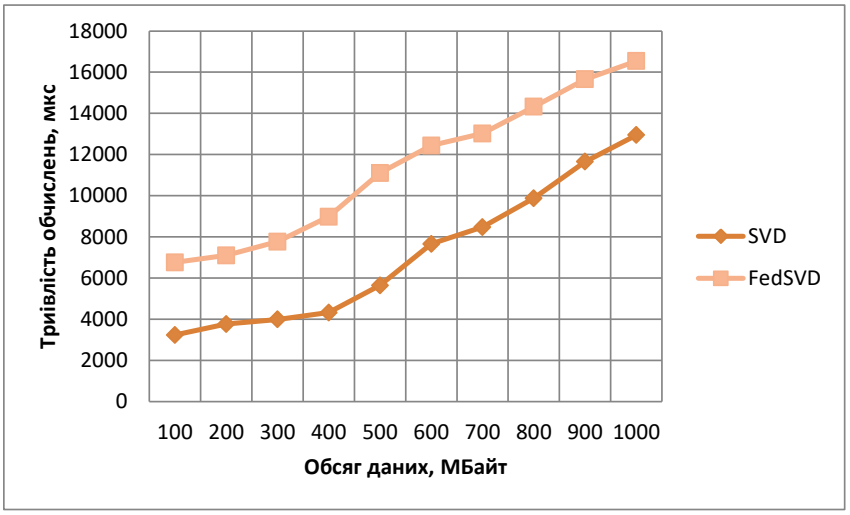

Рис.4. Залежність тривалості обчислень від обсягу обчислюваних даних для розподіленого і нерозподіленого алгоритмів SVD

Як бачимо на рисунку 4, нерозподілений алгоритм сингулярного розкладу працює швидше. Порівняємо тепер точність обчислення сингулярного розкладу для обох алгоритмів (рис.5).

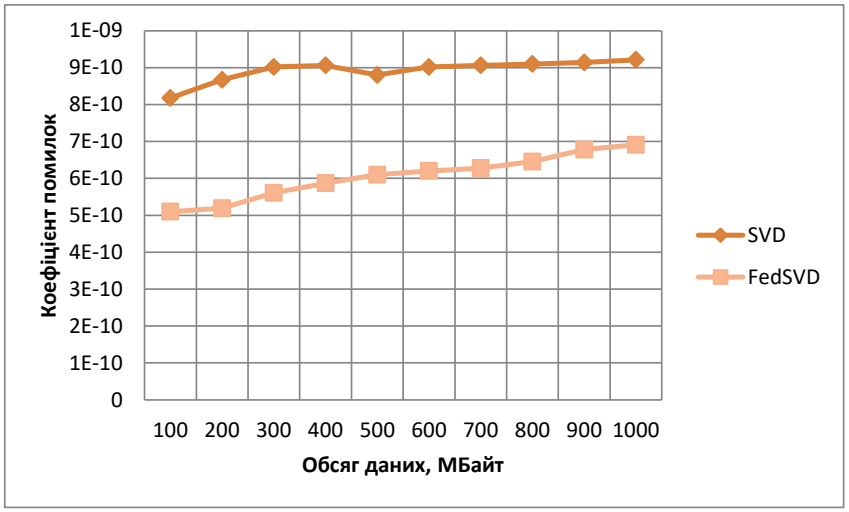

Рис.5. Залежність похибки обчислень від обсягу обчислюваних даних для розподіленого і нерозподіленого алгоритмів SVD

Коефіцієнт помилок Δ показує різницю між початковою матрицею даних та результатом сингулярного розкладу:

$$\Delta = |M - U * \sum * V^T|. \quad (13)$$

Як показує рисунок 5, оскільки в алгоритмі FedSVD використовуються певна кількість менших за розміром під матриць, для кожного з яких обчислюється сингулярний розклад, то точність в такому випадку зростає.

Отримані результати свідчать про переваги алгоритму FedSVD для систем, в яких необхідно забезпечити високу точність та надійність обчислень. Також вони відкривають подальші можливості впровадження даного методу в інфокомунікаційних мережах, промисловому Інтернеті Речей, корпоративних інформаційних системах для покращення ефективності обчислення великих даних.

Висновки. В роботі наведено підхід до оброблення великих обсягів даних в комерційних інформаційних системах за допомогою алгоритму сингулярного розкладу даних. Визначено основні особливості розподіленої обробки масивів інформації. Досліджено математичні особливості алгоритму FedSVD, що дозволяє розподілено обчислювати сингулярний розклад даних. На основі запропонованого методу розроблено програмну модель для оцінки ефективності алгоритму FedSVD, результати роботи якої доводять його високу точність та надійність.

Отримані результати дослідження можуть бути використані при створенні інфокомунікаційних систем, робота яких вимагає швидкої обробки даних, аналізу та передбачень на основі отриманої інформації. Запропонований підхід може слугувати основою для розробки рекомендаційних систем, мереж моніторингу та аналізу даних на великих підприємствах, щоб покращити ефективність та надійність їх роботи .

Список використаних джерел

1. F. Ortega, and A. González-Prieto, “Recommender systems and collaborative filtering,” *Appl. Sci*., vol. 10, PP. 168-173, 2020.

2. Z. Wang, H. Wu, Z. Jiang, P. Ju, J. Yang, Z. Zhou, and X. Chen, “Singular value decomposition-based load indexes for load profiles clustering,” *Transmission Distribution IET Generation*, vol. 14, PP. 4164-4172, 2020.

3. M. Khan, Y. Jin, M. Li, Y. Xiang, and C. Jiang, “Hadoop performance modeling for job estimation and resource provisioning,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, PP. 441–454, 2016.

4. K. Sridharan, G. Komarasamy, S. Daniel Madan Raja, “Hadoop framework for efficient sentiment classification using trees,” *IET Networks*, vol. 9, PP. 223-228, 2020.

5. H. Zhang, Y. Wang, H. Chen, Y. Zhao and J. Zhang, "Exploring machine-learning-based control plane intrusion detection techniques in software defined optical networks," *Optical Fiber Technology*, vol. 39, PP. 37-42, 2017.

6. M. Prakash, G. Singaravel, “Haphazard, enhanced haphazard and personalised anonymisation for privacy preserving data mining on sensitive data sources,” *International Journal of Business Intelligence and Data Mining*, vol. 13, no. 4, PP. 456-474, 2018.

7. M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi and A.P. Sheth, “Machine learning for Internet of Things data analysis: A survey,” *Digital Communications and Networks, Elsevier*, vol.3, PP.34-41, 2017.

8. Т.В. Борис, М.О. Алєксєєв, “Порівняльний аналіз технології паралельного обчислення великих масивів даних MapReduce”, *Second International Conference “Cluster Computing*”, Львів, 2013, С. 1-3.

9. Di Chai, Leye Wang, Lianzhi Fu, Junxue Zhang, Kai Chen, and Qiang Yang, “Federated Singular Vector Decomposition”, *arXiv:2105.08925*, v1, May, 2021.

10. Di Chai, Leye Wang, Kai Chen, and Qiang Yang. “Secure federated matrix Factorization”, *IEEE Intelligent Systems*, 2020.

11. Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth, “Practical secure aggregation for federated learning on user-held data”. *arXiv preprint arXi,* v:1611.04482, 2016.

References

1. F. Ortega, and A. González-Prieto, “Recommender systems and collaborative filtering,” *Appl. Sci*., vol. 10, PP. 168-173, 2020.

2. Z. Wang, H. Wu, Z. Jiang, P. Ju, J. Yang, Z. Zhou, and X. Chen, “Singular value decomposition-based load indexes for load profiles clustering,” *Transmission Distribution IET Generation*, vol. 14, PP. 4164-4172, 2020.

3. M. Khan, Y. Jin, M. Li, Y. Xiang, and C. Jiang, “Hadoop performance modeling for job estimation and resource provisioning,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, PP. 441–454, 2016.

4. K. Sridharan, G. Komarasamy, S. Daniel Madan Raja, “Hadoop framework for efficient sentiment classification using trees,” *IET Networks*, vol. 9, PP. 223-228, 2020.

5. H. Zhang, Y. Wang, H. Chen, Y. Zhao and J. Zhang, "Exploring machine-learning-based control plane intrusion detection techniques in software defined optical networks," *Optical Fiber Technology*, vol. 39, PP. 37-42, 2017.

6. M. Prakash, G. Singaravel, “Haphazard, enhanced haphazard and personalised anonymisation for privacy preserving data mining on sensitive data sources,” *International Journal of Business Intelligence and Data Mining*, vol. 13, no. 4, PP. 456-474, 2018.

7. M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi and A.P. Sheth, “Machine learning for Internet of Things data analysis: A survey,” *Digital Communications and Networks, Elsevier*, vol.3, PP.34-41, 2017.

8. T.V. Boris, M.O. Alekseev, “Comparative analysis of the technology of parallel computation of large data sets MapReduce”, *Second International Conference “Cluster Computing”*, Lviv, 2013, PP. 1-3.

9. Di Chai, Leye Wang, Lianzhi Fu, Junxue Zhang, Kai Chen, and Qiang Yang, “Federated Singular Vector Decomposition”, *arXiv:2105.08925*, v1, May, 2021.

10. Di Chai, Leye Wang, Kai Chen, and Qiang Yang. “Secure federated matrix Factorization”, *IEEE Intelligent Systems*, 2020.

11. Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth, “Practical secure aggregation for federated learning on user-held data”. *arXiv preprint arXi,* v:1611.04482, 2016.