Облікова картка дисертації

I. Загальні відомості

Державний обліковий номер: 0421U102388

Особливі позначки: відкрита

Дата реєстрації: 28-05-2021

Статус: Захищена

Реквізити наказу МОН / наказу закладу:

II. Відомості про здобувача

Власне Прізвище Ім'я По-батькові:

1. Насіров Еміл Мехдієвич

2. Nasirov Emil Mekhdiievych

Кваліфікація:

Ідентифікатор ORCID ID: Не застосовується

Вид дисертації: кандидат наук

Аспірантура/Докторантура: так

Шифр наукової спеціальності: 01.05.01

Назва наукової спеціальності: Теоретичні основи інформатики та кібернетики

Галузь / галузі знань: Не застосовується

Освітньо-наукова програма зі спеціальності: Не застосовується

Дата захисту: 13-05-2021

Спеціальність за освітою: Інформатика

Місце роботи здобувача:

Код за ЄДРПОУ:

Місцезнаходження:

Форма власності:

Сфера управління:

Ідентифікатор ROR: Не застосовується

III. Відомості про організацію, де відбувся захист

Шифр спеціалізованої вченої ради (разової спеціалізованої вченої ради): Д 26.001.09

Повне найменування юридичної особи: Київський національний університет імені Тараса Шевченка

Код за ЄДРПОУ: 02070944

Місцезнаходження: вул. Володимирська, буд. 60, м. Київ, 01033, Україна

Форма власності:

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR: Не застосовується

IV. Відомості про підприємство, установу, організацію, в якій було виконано дисертацію

Повне найменування юридичної особи: Київський національний університет імені Тараса Шевченка

Код за ЄДРПОУ: 02070944

Місцезнаходження: вул. Володимирська, буд. 60, м. Київ, 01033, Україна

Форма власності:

Сфера управління: Міністерство освіти і науки України

Ідентифікатор ROR: Не застосовується

V. Відомості про дисертацію

Мова дисертації:

Коди тематичних рубрик: 16.31.21.07

Тема дисертації:

1. Паралелізація невід'ємної факторизації розріджених лінгвістичних матриць та тензорів надвеликої розмірності

2. Parallelization of non-negative huge sparse linguistic matrix and tensors factorization

Реферат:

1. В роботі запропоновані паралельні методи невід'ємної факторизації надвеликих розріджених матриць та тензорів - популярний метод в комп'ютерній лінгвістиці. Проблема невід'ємної факторизації розріджених матриць постала в процесі розробки системи визначення міри семантичної близькості-зв'язності за технологією Латентного Семантичного Аналізу. Існуючі паралельні моделі для невід'ємної факторизації матриць та тензорів не задовольняють потреби розмірності матриці та тензору або вимагають занадто великих обчислювальних ресурсів. Запропоновано дві методи паралелізації алгоритму факторизації матриць: локальна алгоритм з використанням жорсткого диску та GPU і розподілена модель з використанням мережі вузлів та використання GPU. Ітеративні правила оновлення були розділені на кроки для досягнення мінімальної кількості обчислень таким чином, щоб знизити кількість операцій надлишкового

копіювання пам'яті та мережевих операцій передачі даних. Були співставленні три моделі розподілу алгоритму факторизації матриць. Використання пам'яті і об’єми передачі даних необхідні для роботи алгоритму факторизації були проаналізовані та оптимізовані. Описані локальна модель з використанням GPU та розподілена модель були реалізовані, випробувані та порівнянні в розумінні об'ємів читання та запису на жорсткий диск та передачі по мережі вузлів. Також проаналізовано та порівняно час необхідний для виконання ітерації. В роботі запропоновано блочно-діагональний підхід до факторизації невід'ємних розріджених лінгвістичних матриць, які можуть бути приведені до блочно-діагональної форми. Цей підхід може прискорити факторизацію, потребує менше мережевих операцій та пам'яті для ітерацій і зберігання результатів. Основною ідеєю алгоритму приведення лінгвістичного тензору до блочно-діагонального виду є групування слів однієї тематичної групи разом по всіх осях у відповідних інтервалах для підгонки всіх ненульових значень лінгвістичного тензору на перетині всередину блоку, що складається в даний момент. Вся суть методу полягає у переході від необхідності факторизувати надвеликий розріджений лінгвістичний тензор до невід’ємної факторизації набору лінгвістичних тензорів значно зменшеного розміру. Вказано, що не кожна матриця або тензор можуть бути зведені до блочно-діагональної форми використовуючи перестановки рядків та стовпчиків в матрицях та шарів в тензорах. У випадку лінгвістичних матриць та тензорів допускається розщеплення векторів семантико-синтаксичної валентності слів на складові вектори їх окремих значень. Запропоновано використання особливостей природньої мови присутніх в лінгвістичних матрицях та тензорах для зведення до блочно-діагональної форми, а саме, виділення тематичних діагональних блоків матриць. Запропоновано використання латентного розподілу Діріхле для приведення матриць і тензорів до блочно-діагональної формі для паралелізація обчислень та прискорення невід’ємної факторизації лінгвістичних матриць і тензорів надвеликої розмірності. Запропонований метод, так само, дозволяє доповнення моделі природної мови новими даними без необхідності виконувати невід’ємну факторизацію всього надвеликого тензора заново з самого початку.

2. The paper describes algorithms and methods for parallelizing non-negative factorization of sparse matrices and tensors - popular technology in artificial intelligence in general, and in computational linguistics in particular. Two methods of parallelization of the algorithm for factorization of non-negative matrices are proposed: a local algorithm using a hard disk and computations on GPUs and a distributed algorithm using a network of computational nodes and GPUs. The paper also proposes a block-diagonal approach to factorization of inherent sparse linguistic matrices and tensors, which can be reduced to a block-diagonal form. The proposed method also allows the model to be supplemented with new data without no need to perform the nonnegative factorization of the entire super-large tensor from the very beginning. It is also proposed to use the latent Dirichlet distribution to reduce matrices and tensors to the block-diagonal form by constructing thematic diagonal blocks.

Державний реєстраційний номер ДіР:

Пріоритетний напрям розвитку науки і техніки:

Стратегічний пріоритетний напрям інноваційної діяльності:

Підсумки дослідження:

Публікації:

Наукова (науково-технічна) продукція:

Соціально-економічна спрямованість:

Охоронні документи на ОПІВ:

Впровадження результатів дисертації:

Зв'язок з науковими темами:

VI. Відомості про наукового керівника/керівників (консультанта)

Власне Прізвище Ім'я По-батькові:

1. Марченко Олександр Олександрович

2. Marchenko Oleksandr Oleksandrovich

Кваліфікація: д. ф.-м. н., 01.05.03

Ідентифікатор ORCID ID: Не застосовується

Додаткова інформація:

Повне найменування юридичної особи:

Код за ЄДРПОУ:

Місцезнаходження:

Форма власності:

Сфера управління:

Ідентифікатор ROR: Не застосовується

VII. Відомості про офіційних опонентів та рецензентів

Офіційні опоненти

Власне Прізвище Ім'я По-батькові:

1. Жежерун О. П.

2. Zhezherun O. P.

Кваліфікація: к. ф.-м. н., 01.05.03

Ідентифікатор ORCID ID: Не застосовується

Додаткова інформація:

Повне найменування юридичної особи:

Код за ЄДРПОУ:

Місцезнаходження:

Форма власності:

Сфера управління:

Ідентифікатор ROR: Не застосовується

Власне Прізвище Ім'я По-батькові:

1. Дорошенко Анатолій Юхимович

2. Doroshenko Anatolii Yukhymovych

Кваліфікація: д.ф.-м.н., 01.05.03

Ідентифікатор ORCID ID: Не застосовується

Додаткова інформація:

Повне найменування юридичної особи:

Код за ЄДРПОУ:

Місцезнаходження:

Форма власності:

Сфера управління:

Ідентифікатор ROR: Не застосовується

Рецензенти

VIII. Заключні відомості

Власне Прізвище Ім'я По-батькові голови ради Анісімов Анатолій Васильович

Власне Прізвище Ім'я По-батькові головуючого на засіданні Анісімов Анатолій Васильович

Відповідальний за підготовку облікових документів

Реєстратор

Керівник відділу УкрІНТЕІ, що є відповідальним за реєстрацію наукової діяльності Юрченко Т.А.