РЕЦЕНЗІЯ

доктора технічних наук, професора,

професора кафедри комп'ютерних наук

Чернівецького національного університету імені Юрія Федьковича

УГРИНА ДМИТРА ІЛЛІЧА

на дисертаційне дослідження

Кириченка Євгена Олександровича

на тему «Оптимізація структури гетерогенних даних в Big Data»,

представлену на здобуття наукового ступеня доктора філософії

за спеціальністю 121 – Інженерія програмного забезпечення

галузі знань 12 – Інформаційні технології

Актуальність теми дисертації.

Актуальність теми дисертаційної роботи зумовлена стрімким зростанням обсягів гетерогенних даних, які генеруються сучасними корпоративними та хмарними інформаційними системами з різнорідних і, як правило, слабо документованих джерел. За таких умов класичні підходи до порівняння та структурування табличних наборів даних, що ґрунтуються на повному перенесенні й скануванні копій даних, стають економічно та технологічно неефективними: вони потребують значних обчислювальних ресурсів, створюють надлишкове навантаження на мережу та сховища, а також суттєво ускладнюють інтеграцію систем аналізу у великих корпусах даних. Це підкреслює гостру потребу у розробці принципово нових, масштабованих методів, орієнтованих на компактне представлення даних без втрати інформативних структурних характеристик.

Особливої ваги в сучасних умовах набувають задачі, що виникають на перетині інженерії програмного забезпечення, аналізу великих даних та хмарних технологій: автоматична типізація змінних у слабо структурованих наборах,

побудова уніфікованих метрик структурної подібності для різнотипних змінних, конструювання графових та ієрархічних моделей подібності корпусів даних, а також проєктування масштабованих програмних архітектур, здатних ефективно функціонувати у розподілених хмарних середовищах. Саме на вирішення цих завдань спрямована дисертаційна робота Кириченка Є.О., що й визначає її високу актуальність.

Практична значущість теми обумовлена широким спектром застосування отриманих результатів у реальних прикладних задачах: оптимізації корпоративних сховищ даних, побудові систем виявлення дублікатів та структурних змін, підтримці процесів каталогізації та дослідження корпусів табличних даних, а також у сучасних платформах штучного інтелекту та машинного навчання, де якість і компактність представлень вхідних даних безпосередньо визначає ефективність подальших моделей. З огляду на це, тема дослідження є безсумнівно актуальною як для наукової спільноти у галузі інформаційних технологій, так і для практичного застосування.

Зв'язок роботи з науковими програмами, планами, темами.

Наукові дослідження були виконані здобувачем на кафедрі програмного забезпечення комп'ютерних систем та кафедрі математичних проблем управління і кібернетики Чернівецького національного університету імені Юрія Федьковича в межах держбюджетних науково-дослідних робіт «Дослідження, моделювання та розробка програмного забезпечення складних динамічних систем» (№ 0121U109232) та «Інформаційні технології в аспекті сучасних задач прийняття рішень» (№ 0121U109159).

Ступінь обґрунтованості наукових положень, висновків, рекомендацій, сформульованих у дисертації.

Наукові положення, які представлені в дисертаційній роботі, є добре обґрунтованими, а також належно висвітлені у відповідних розділах дисертації. Основні результати, отримані здобувачем та винесені на захист, цілком

відповідають меті та завданням роботи, обговорювалися на наукових семінарах кафедри програмного забезпечення комп'ютерних систем та кафедри математичних проблем управління і кібернетики ННІФТКН, наукових конференціях та опубліковані у фахових виданнях та у наукових журналах, які індексуються в базах даних Scopus. Достовірність отриманих результатів ґрунтується на використанні загальноприйнятих експериментальних і теоретичних підходів й методів дослідження та не викликає сумнівів. У роботі проведено ґрунтовний огляд літературних джерел з тематики дисертаційного дослідження.

Наукова новизна

Наукова новизна результатів дисертаційного дослідження полягає в наступному:

1. Вперше запропоновано уніфіковану типо-орієнтовану методологію формування компактних дескрипторів для різнотипних атрибутів табличних наборів даних. На відміну від відомих рішень, що зазвичай розраховані на заздалегідь визначену схему, цей підхід забезпечує побудову інтерпретованих зведень для різних типів змінних: таблиць частот для факторних, гістограм для часових, векторів моментів до четвертого порядку для числових та TF-IDF векторів для рядкових. Обґрунтованість запропонованого підходу підтверджується математичним апаратом, представленим у підрозділах 2.2 та 2.3, а достовірність – результатами експериментального тестування на реальних фінансових часових рядах (розділ 4), де показано, що застосування типо-орієнтованих зведень суттєво зменшує обсяг даних для зберігання й передачі при збереженні інформативної спроможності, достатньої для подальшого порівняння наборів даних.

2. Запропоновано нову зважену метрику структурної подібності даних DISS, яка здійснює агрегацію відстаней Геллінґера, Вассерштейна ($p = 1$), повної варіації та L1/L2-норм у поєднанні з MAE/MAPE для числових зведень, що забезпечує принципово обґрунтоване порівняння характеристик різнотипних змінних у межах

єдиної метричної моделі. Наукова новизна полягає в тому, що, на відміну від підходів, які використовують одиничні метрики або векторні представлення, DISS-метрика враховує специфіку типів змінних через систему ваг, що можуть бути як встановлені експертом, так і отримані в результаті навчання. Коректність і обґрунтованість метрики детально розглянуто у підрозділах 2.4.1–2.4.3, де наведено як математичні моделі, так і алгоритмічні принципи побудови DISS-матриць та виявлення зсувів розподілів.

3. Вперше запропоновано метод формування ієрархічних структур для оцінки близькості корпусів табличних наборів, який ґрунтується на графі суміжності та дереві similarity tree, сконструйованому з використанням алгоритму мінімального остовного дерева у метричному просторі DISS. Такий підхід дозволяє виявляти кластерну організацію корпусів без попередньої інформації щодо схеми даних та підтримує багаторівневе ієрархічне дослідження наборів. Експериментальні дослідження (підрозділи 4.3–4.5) свідчать про ефективність такого підходу при пошуку найбільш подібних даних, ідентифікації ключових зв'язків у наборах та виявленні структурних відхилень, що становить вагомий крок у розвитку графово-спектральних методів кластеризації великих гетерогенних корпусів.

4. Вперше спроєктовано та реалізовано наскрізну інформаційну технологію опрацювання великих колекцій табличних даних у хмарному середовищі з підтримкою горизонтального масштабування. На відміну від класичних рішень, які оперують повними копіями даних, тут обчислення компактних дескрипторів, їх збереження та виконання запитів розподіляються між сервісами AWS S3, Glue, Athena, Airflow та обчислювальним ядром Apache Spark. Запропонована у розділі 3 архітектура дає змогу суттєво зменшити витрати на передачу й збереження масивів даних, не погіршуючи якості порівняння наборів, а експериментальні вимірювання, представлені у розділі 4, фіксують зростання продуктивності обчислень у межах 40–60 % відносно сценаріїв повного сканування таблиць.

Оригінальність рішення полягає в органічному поєднанні теоретично обґрунтованої моделі CDR+DISS із хмарною платформою, що робить отримані результати придатними для прямого впровадження у сучасні інформаційні системи.

5. Дістав подальшого розвитку підхід до кластеризації різнотипних даних, побудований на поєднанні компактних дескрипторів із матрицею суміжності, що надає алгоритму властивостей шумостійкості та масштабованості. Автором запропоновано досліджувати кластерну будову через аналіз спектральних характеристик стохастичних матриць і обґрунтовано критерій підбору числа кластерів, що спирається на кількість власних значень, які перетинають встановлений пороговий рівень (розділ 1). Розроблений інструментарій придатний для автоматизованого виділення природних груп ознак та ідентифікації аномальних структурних змін у великих сукупностях даних, розширюючи таким чином набір графово-спектральних методів кластеризації у задачах Big Data.

Достовірність отриманих наукових результатів підкріплена коректним застосуванням математичного апарату теорії ймовірностей, математичної статистики, теорії випадкових матриць, а також методів спектрального аналізу лінійної алгебри. Результати експериментальної верифікації, проведеної на фінансових часових рядах із двох відкритих джерел, повністю узгоджуються з теоретичними висновками автора, що підтверджує наукову цінність роботи.

Практичне значення одержаних результатів.

Запропонований підхід до побудови компактних представлень (CDR) та метрики DISS для порівняння табличних даних та розроблений метод кластеризації на основі структурних відстаней використовуються у роботі компаній ТОВ «Кодерс ПРО» та ТОВ «Палетний сервіс». А результати теоретичних та практичних досліджень впроваджено у навчальний процес кафедр математичних проблем управління і кібернетики та програмного забезпечення комп'ютерних систем Чернівецького національного університету імені Юрія

Федьковича.

Оцінка змісту дисертації, її завершеність та дотримання принципів академічної доброчесності.

За своїм змістом дисертаційна робота здобувача Кириченка Є.О. повністю відповідає Стандарту вищої освіти зі спеціальності 121 – Інженерія програмного забезпечення та напрямам досліджень відповідно до освітньої програми «Інженерія програмного забезпечення».

Дисертаційна робота є завершеною науковою працею і свідчить про наявність особистого внеску здобувача у науковий напрям інженерії програмного забезпечення, зокрема у створенні програмних засобів аналізу та структурування гетерогенних табличних даних у хмарному середовищі.

Розглянувши звіт подібності за результатами перевірки дисертаційної роботи на текстові співпадіння, можна зробити висновок, що дисертаційна робота Кириченка Євгена Олександровича є результатом самостійних досліджень здобувача і не містить елементів фальсифікації, компіляції, фабрикації, плагіату та запозичень. Використані ідеї, результати і тексти інших авторів мають належні посилання на відповідне джерело.

Мова та стиль викладення результатів.

Дисертаційна робота написана українською мовою.

Робота характеризується належним рівнем наукового викладу та логічною структурою подання матеріалу: спочатку наведено теоретичні засади аналізу великих гетерогенних даних із застосуванням спектральних методів та апарату випадкових матриць, далі – авторські математичні та алгоритмічні розробки компактного представлення і метрик подібності, потім – архітектурні рішення щодо хмарної інформаційної системи та, нарешті, експериментальна перевірка ефективності запропонованих підходів. Мова викладення є грамотною, технічно точною, з доречним використанням термінів, характерних для інженерії програмного забезпечення, аналізу Big Data та хмарних обчислень. Теоретичні

підрозділи супроводжуються обґрунтованими схемами, формулами і прикладами, що суттєво підвищує доступність викладеного матеріалу. Робота легко сприймається, незважаючи на складність тематики.

Дисертація складається зі вступу, 4 розділів, висновків, списку використаних джерел та додатків. Загальний обсяг дисертації становить 224 сторінки, з яких 145 сторінок основного тексту, 27 сторінок – список використаних джерел та 34 сторінки – додатки.

У *вступі* автор обґрунтовує актуальність обраної теми в контексті сучасного стану аналізу великих гетерогенних даних та хмарних інформаційних систем. Чітко сформульовано об'єкт, предмет, мету і завдання дослідження, обрані методи та інструменти, а також викладено елементи наукової новизни й практичного значення.

Перший розділ має оглядово-аналітичну та теоретичну спрямованість. Автор системно опрацьовує спектральні методи аналізу кластерної структури великих даних у гетерогенних мережевих середовищах, представляє математичний апарат стохастичних і випадкових матриць, формулює гіпотези кластерної будови графу та обґрунтований критерій вибору оптимального числа кластерів, а також розглядає прикладні задачі Big Data, на яких розкривається важливість запропонованого методу.

Другий розділ присвячений математичним основам та алгоритмам компактного представлення даних. Автор послідовно розкриває задачу визначення подібності, формулює вимоги до компактних представлень, описує алгоритм автоматичної типізації змінних, підходи до стиснення різних типів змінних та будує метрику DISS із детальним розглядом її компонентів, а також формулює принципи побудови DISS-матриць і алгоритми виявлення дублікатів та зсувів розподілів.

Третій розділ має прикладну спрямованість і присвячений проєктуванню інформаційної системи для оптимізації структури гетерогенних даних. Розглянуто

функціональне призначення системи, вимоги до неї, логічну архітектуру та архітектуру хмарної реалізації на базі AWS (S3, EMR Serverless, Glue, Athena, Airflow), а також технологічний стек і принципи масштабування. Описано функціонування системи, послідовність обробки даних та інтеграційні сценарії.

Четвертий розділ охоплює реалізацію інформаційної системи та експериментальну верифікацію запропонованих підходів. Апробацію здійснено на фінансових часових рядах, взятих із двох відкритих джерел. Проведено оцінку структурної стабільності дескрипторів при поетапному скороченні обсягу вхідних даних, аналіз чутливості компактних представлень, дослідження кривої DISS-дрейфу, ROC-оцінку якості виявлення змін та порівняння ступеня схожості ознак, а також виконано ідентифікацію ключових зв'язків у наборах. Зіставлення з базовими методами наочно показує переваги авторського рішення.

У висновках підсумовано результати дослідження; додатки містять акт впровадження, список опублікованих праць автора, лістинг програмного забезпечення.

Оприлюднення результатів дисертаційної роботи.

Наукові результати дисертації висвітлені у 13 наукових публікаціях здобувача, серед яких: 2 статті у виданнях, у наукометричній базі Scopus, 4 статті у наукових виданнях, включених до переліку наукових фахових видань України, з них 1 публікація, що додатково відображає наукові результати дисертації, а також 6 публікацій у матеріалах міжнародних та всеукраїнських науково-практичних конференцій.

Результати дисертації були апробовані на 6 наукових конференціях, серед яких міжнародні: «Проблеми інформатики та комп'ютерної техніки (ПІКТ)» (Чернівці, 2023–2025), «Інформаційні технології: наука, техніка, технологія, освіта, здоров'я (MicroCAD-2024)» (Харків), The 13th International Conference on Electronics, Communications and Computing's (IC ECCO) (Кишинів, Молдова, 2024),

«Science in the Context of Modern Challenges and Prospects» (Утрехт, Нідерланди, 2025).

Наукові результати, описані в дисертаційній роботі, повністю висвітлені у наукових публікаціях здобувача. В усіх публікаціях дотримано принципів академічної доброчесності. Особистий внесок автора у спільних роботах чітко простежується та повністю відповідає результатам, зарахованим за темою дисертаційного дослідження.

Недоліки та зауваження до дисертаційної роботи.

1. У розділі 1, присвяченому аналізу спектральних методів кластеризації, наведено ґрунтовний теоретичний огляд сучасних підходів, однак відсутня чітка систематизація обмежень існуючих методів у контексті роботи з гетерогенними табличними даними. Зокрема, доцільно було б узагальнити недоліки розглянутих підходів у вигляді порівняльної таблиці або структурованого висновку, що дозволило б чіткіше обґрунтувати необхідність запропонованого автором методу.

2. У підрозділі 2.3 описано процедуру стиснення числових змінних через вектори моментів до четвертого порядку, проте не наведено кількісного порівняння інформаційних втрат у залежності від порядку моменту, що ускладнює вибір оптимальної глибини зведення для прикладних задач з високою варіабельністю розподілів.

3. У підрозділі 2.4.1 під час обґрунтування ваг у DISS-метриці подано загальні рекомендації щодо їх встановлення експертом або навчання, проте відсутній систематичний аналіз чутливості метрики до вибору ваг для різних класів корпусів даних, що могло б посилити узагальненість висновків.

4. Розділ 3 містить опис архітектури системи на базі AWS (S3, EMR Serverless, Glue, Athena, Airflow), однак характеризується переважно якісним описом переваг кожного компонента. Не наведено кількісних показників: залежності часу обробки від обсягу даних, залежності продуктивності від кількості executor'ів у Spark-кластері, обсягу використаної оперативної пам'яті,

часу cold-start ініціалізації EMR-кластера. Доцільно було б провести експеримент з нарощуванням обсягу даних принаймні до десятків ГБ та виміряти показники прискорення обчислень.

5. Експериментальні дослідження у розділі 4 виконано на фінансових часових рядах із двох відкритих наборів даних, що доцільно було розширити за рахунок корпусів з інших предметних областей (наприклад, біомедичних або сенсорних даних IoT) для підтвердження універсальності запропонованого підходу.

6. У роботі недостатньо уваги приділено аналізу обчислювальної складності запропонованих алгоритмів, зокрема оцінці їх масштабованості при збільшенні кількості ознак та об'єктів у наборі даних. Включення асимптотичних оцінок складності та експериментального підтвердження ефективності алгоритмів при великих обсягах даних дозволило б підсилити практичну значимість дослідження.

7. У дисертаційній роботі спостерігається певна нерівномірність між теоретичною та експериментальною частинами: при достатньо глибокому математичному обґрунтуванні моделей експериментальна валідація могла б бути більш розширеною та різноплановою. Зокрема, додаткове тестування на різнорідних типах даних та детальніший аналіз отриманих результатів підвищили б рівень доказовості зроблених висновків.

Вважаю, що висловлені зауваження не є визначальними і не зменшують загальну наукову новизну та практичну значимість результатів та не впливають на позитивну оцінку дисертаційної роботи.

Висновок про дисертаційну роботу.

Вважаю, що дисертаційна робота здобувача ступеня доктора філософії Кириченка Євгена Олександровича на тему «Оптимізація структури гетерогенних даних в Big Data» виконана на належному науковому рівні, не порушує принципів академічної доброчесності та є закінченим науковим дослідженням, сукупність теоретичних та практичних результатів якого розв'язує наукове завдання, що має

істотне значення для інформаційних технологій. Дисертаційна робота за актуальністю, практичною цінністю та науковою новизною повністю відповідає вимогам чинного законодавства України, що передбачені в п. 6, 7, 8, 9 «Порядку присудження ступеня доктора філософії та скасування рішення разової спеціалізованої вченої ради закладу вищої освіти, наукової установи про присудження ступеня доктора філософії», затвердженого Постановою Кабінету Міністрів України від 12 січня 2022 р. № 44.

Здобувач Кириченко Євген Олександрович заслуговує на присудження ступеня доктора філософії в галузі знань 12 – Інформаційні технології за спеціальністю 121 – Інженерія програмного забезпечення.

Рецензент:
професор кафедри комп'ютерних наук
Чернівецького національного університету
імені Юрія Федьковича,
доктор технічних наук, професор Дмитро УГРИН

Підпис Угрина Д. засвідчую
Учений секретар Чернівецького національного університету імені Юрія Федьковича
Мирославська А.
01 травня 26