ВІДГУК

офіційного опонента – доктора технічних наук, професора,
професора кафедри комп'ютерних систем, мереж та кібербезпеки
Національного університету біоресурсів і природокористування України
КРИВОРУЧКО Олени Володимирівни
на дисертаційне дослідження **КИРИЧЕНКА Євгена Олександровича**
на тему: **«Оптимізація структури гетерогенних даних в Big Data»,**
подану на здобуття наукового ступеня доктора філософії
за спеціальністю 121 – Інженерія програмного забезпечення
галузі знань 12 – Інформаційні технології

Актуальність дисертаційного дослідження.

Сьогоднішній етап розвитку інформаційних технологій характеризується, безумовно, тим, що людство оперує величезними обсягами різноманітної гетерогенної інформації, для чого найчастіше використовують методи інтелектуального аналізу даних та машинного навчання. Більше того, інформаційний простір, в якому ми живемо і працюємо, часто являє собою складну гетерогенну структуру з табличними наборами даних із різних, часто слабо документованих джерел, яка потребує дослідження. Велике значення при цьому має адекватна сегментація та компактне подання наборів даних, оскільки аналіз повнорозмірних даних є дуже ресурсозатратним. Кластеризація та побудова графів подібності дозволяють спростити такий аналіз, дослідити структуру корпусу даних в цілому та його частин зокрема.

Не дивлячись на те, що процес інтелектуального аналізу великих даних достатньо забезпечений математичним, алгоритмічним та програмним забезпеченням, існує ряд проблем, що потребує вирішення. Зокрема, це побудова компактних представлень для гетерогенних табличних даних без необхідності оперувати повними копіями інформації, створення універсальних метрик для порівняння різнорідних структур даних, визначення оптимальної кількості кластерів на основі спектральних методів,

а також реалізація масштабованої хмарної інформаційної технології для обробки великих корпусів даних.

Дисертаційна робота Кириченка Євгена Олександровича є вдалою спробою вирішити зазначені проблеми, і, зважаючи на наведене вище, вибір тематики та дослідження і поставлені перед дисертантом завдання є, безумовно, актуальними.

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

Дисертаційну роботу Кириченка Євгена Олександровича виконано на кафедрі програмного забезпечення комп'ютерних систем Чернівецького національного університету імені Юрія Федьковича. Її зміст відповідає тематиці науково-дослідних робіт: «Дослідження, моделювання та розробка програмного забезпечення складних динамічних систем» (Державний реєстраційний номер 0121U109232) та «Інформаційні технології в аспекті сучасних задач прийняття рішень» (Державний реєстраційний номер 0121U109159). Варто зазначити, що кафедра провадить діяльність за акредитованою аспірантською програмою «Інженерія програмного забезпечення», яку закінчив дисертант. Тематика рецензованої дисертаційної роботи повністю відповідає зазначеній освітньо-науковій програмі III рівня вищої освіти та вище зазначеним темам науково-дослідних робіт.

Структура дисертації.

Дисертація є завершеною науково-дослідною роботою, що складається з анотації (українською та англійською мовами), змісту, переліку умовних скорочень, чотирьох розділів, висновків, списку використаних джерел та додатків.

У вступі обґрунтовано актуальність теми дисертації, визначено об'єкт та предмет дослідження, сформульовано мету та задачі, наукову новизну та практичне значення отриманих результатів, подано відомості про апробацію роботи та наведено її структуру.

Перший розділ роботи присвячено спектральним підходам до кластеризації у Big Data та гетерогенних мережевих системах. Дисертант

розглянув роль випадкових матриць у моделюванні складних мереж, побудував математичну модель на основі матриці суміжності з незалежними однаково розподіленими елементами, дослідив стохастичні моделі графів великої розмірності. Ключовим результатом є формулювання гіпотез кластерної структури графу та обґрунтування критерію визначення оптимальної кількості кластерів на основі підрахунку власних значень стохастичної матриці, які перевищують заданий поріг. Матеріал розділу висвітлено повно, проведено ґрунтовний аналіз предметної галузі, ретельно та всебічно обґрунтовано актуальність та завдання дисертаційного дослідження.

У другому розділі дисертант розробив математичний та алгоритмічний апарат інформаційної технології компактного подання, класифікації та порівняння табличних даних. Запропоновано узагальнену модель компактного представлення (CDR), що замінює повні набори даних стислими дескрипторами, які зберігають інформаційно значущі статистичні та структурні характеристики. Розроблено алгоритм автоматичної типізації змінних, описано класи метрик для числових, текстових та категоріальних даних, побудовано та обґрунтовано конвеєрну систему аналізу даних. Особливу увагу заслуговує запропонована метрика структурної подібності DISS, яка агрегує відстані Геллінґера, Вассерштейна, повної варіації та L1/L2.

У третьому розділі здійснено аналіз сучасних хмарних технологій та розроблено багаторівневу модульну архітектуру інформаційної системи. Програмне забезпечення реалізовано мовою Python із використанням Apache Airflow для оркестрації та Apache Spark для розподіленої обробки даних. Описано повний конвеєр трансформації неструктурованих даних із застосуванням AWS-сервісів (S3, EMR, Glue, Athena). Варто зазначити, що архітектура забезпечує незалежність обчислювального ядра від конкретного постачальника хмарних послуг.

Четвертий розділ дисертації присвячено реалізації інформаційної системи та експериментальній перевірці інформаційної технології моделювання та структурного аналізу гетерогенних табличних даних. Експериментальне тестування проводилось на реальних фінансових часових рядах із двох відкритих наборів даних. Отримані результати демонструють, що система CDR/DISS дозволяє точно виявляти зміни в структурі даних навіть при значному зменшенні обсягів інформації, а застосування розподіленої архітектури забезпечило прискорення обробки на 40–60% порівняно з традиційними підходами.

Висновки дисертаційної роботи сформульовано чітко, вони повністю охоплюють отримані результати і повністю задовольняють вимоги, що ставляться до результатів дисертації на здобуття наукового ступеня доктора філософії.

Список використаних джерел повно охоплює предметну галузь і вказує на аналіз значної кількості літературних джерел, як українських, так і закордонних авторів.

Додатки містять список публікацій за темою дисертації, відомості про апробацію, акти впровадження результатів дисертаційної роботи та лістинг частини коду програмного забезпечення.

Наукова новизна одержаних результатів.

До найсуттєвіших та нових наукових результатів дисертаційної роботи Кириченка Є. О., на мою думку, можна віднести наступне:

1. Автором вперше:

– розроблено уніфіковану методологію конструювання компактних представлень даних (CDR) для гетерогенних табличних змінних, яка, на відміну від існуючих підходів, забезпечує типо-орієнтовану побудову інтерпретованих зведень через таблиці частот для факторних змінних, гістограми для часових змінних, вектори моментів до четвертого порядку для числових змінних та TF-IDF вектори для рядкових змінних;

– запропоновано нову метрику структурної подібності даних (DISS), що здійснює зважену агрегацію відстаней Геллінгера, Вассерштейна, повної варіації та L1/L2 з MAE/MAPE для числових зведень для забезпечення принципово обґрунтованого порівняння характеристик різнотипних змінних;

– розроблено метод побудови ієрархічних структур подібності корпусів табличних даних на основі графів суміжності, який дозволяє пошук найбільш подібних даних та ієрархічне дослідження без припущення фіксованої схеми;

– розроблено та реалізовано наскрізну хмарну масштабовану інформаційну технологію обробки великих корпусів табличних даних на базі Apache Spark, AWS S3, Glue, Athena та Airflow.

2. Набуло подальшого розвитку:

– теорія структурної подібності табличних даних шляхом узагальнення статистичних відстаней для різних типів змінних;

– графово-спектральні підходи до кластеризації, у яких подібність між наборами даних задається компактними дескрипторами, а кластерна структура визначається через спектр матриці подібності.

Вважаю, що отримані дисертантом наукові результати є вагомим внеском у методи та засоби дослідження та порівняння гетерогенних даних за допомогою інтелектуального аналізу даних.

Достовірність отриманих результатів та висновків.

Достовірність результатів, отриманих в дисертаційній роботі Кириченка Є. О., забезпечено коректною постановкою задачі, мети та завдань дослідження, послідовним їх розв'язанням та аргументованим вибором методів та програмних засобів розробки.

Достовірність наукових положень, висновків та рекомендацій підтверджуються сучасними методами досліджень, які відповідають поставленій задачі, глибоким аналізом об'єкта та предмета дослідження за допомогою адекватних методів, а також підтверджується публікаціями основних результатів дисертаційної роботи та успішною апробацією на всеукраїнських та міжнародних науково-практичних конференціях.

Практична цінність одержаних результатів.

Практична цінність наукових результатів дисертаційної роботи Кириченка Є. О. полягає у практичному застосуванні теоретичних положень, методів та технологій, що підтверджується актами про впровадження результатів дисертації. Зокрема, запропонований підхід до побудови компактних представлень (CDR) та метрики DISS для порівняння табличних даних, розроблений метод кластеризації на основі структурних відстаней та методи інтеграції кластерного аналізу використовуються у роботі компаній ТОВ «Кодерс ПРО» та ТОВ «Палетний сервіс».

Варто зауважити, що результати дисертаційного дослідження використовуються у навчальному процесі кафедр математичних проблем управління і кібернетики та програмного забезпечення комп'ютерних систем Чернівецького національного університету імені Юрія Федьковича.

Оформлення результатів, дотримання вимог академічної доброчесності, повнота викладу наукових положень та результатів у публікаціях.

Дисертація має повний обсяг 224 сторінки друкованого тексту, причому основна частина викладена на 145 сторінках. Список використаних джерел є репрезентативним.

Дисертаційна робота має логічну структуру, висновки та рекомендації відповідають отриманим у розділах результатам. Оформлення дисертації задовольняє усі вимоги до такого роду кваліфікаційних наукових праць.

Результати перевірки дисертаційної роботи на наявність академічного плагіату свідчать про високу індивідуальність роботи. Авторський стиль простежується по всьому тексту дисертації. Відсутні запозичення і використання результатів інших авторів без посилання на їхні джерела. Це підтверджує, що дисертаційна робота Кириченка Є. О. відповідає нормам академічної доброчесності.

Основні положення дисертації та найважливіші її результати опубліковано у науковій періодиці та апробовані на наукових конференціях.

Зокрема, дисертантом опубліковано 2 статті у наукових виданнях, які індексуються у наукометричній базі SCOPUS; 4 статті – у виданнях, включених до переліку наукових фахових видань України; 6 робіт – у збірниках матеріалів міжнародних та всеукраїнських наукових конференцій. Отже, вимоги щодо кількості та якості наукових публікацій автором виконано.

Дискусійні положення та зауваження до змісту дисертаційного дослідження.

Однак, дисертаційна робота Кириченка Є. О. не позбавлена деяких недоліків. До них можна віднести наступне:

1. При викладенні матеріалу у розділі 3 пункт 3.2 (стор. 94-95) бажано було б приділити більше уваги опису нефункціональних вимог до розробленої інформаційної системи, зокрема питанням безпеки, надійності та зручності використання (usability).

2. Обґрунтування хмарної реалізації передбачає, серед іншого, і зниження витрат на зберігання та обробку даних, однак у роботі відсутня розгорнута економічна модель, яка б формально пов'язувала обсяг CDR-представлень, кількість запусків конвеєра та режими роботи EMR з відповідними показниками сукупних експлуатаційних витрат. Доповнення експериментальної частини таким аналізом (наприклад, порівнянням вартості на 1 ГБ вхідних даних у режимах EMR на EC2 та EMR Serverless) надало б отриманим інженерним рішенням додаткового економічного обґрунтування.

3. Архітектура (підрозділ 3.3) жорстко прив'язана до екосистеми AWS: S3, Glue, Athena, EMR, EMR Serverless, Airflow на MWAA. Така vendor-lock-in архітектура суперечить заявленим у вступі принципам переносимості та інтероперабельності інформаційної системи. Бажано було сформулювати архітектуру у термінах абстракцій (object storage, metadata catalog, distributed compute engine, workflow orchestrator) і лише потім зіставити з конкретними сервісами AWS, з паралельними аналогами в Azure (Blob Storage, Synapse,

Data Factory) та GCP (Cloud Storage, Dataproc, Composer). У такій формі робота мала б суттєво вищу практичну цінність та довговічність.

4. У таблиці 4.5 наведено зменшення розміру даних «більш ніж на 99,99% без втрати статистично значущої інформації». Твердження потребує строгого доведення: яка саме статистично значуща інформація зберігається? Запропонований CDR за побудовою втрачає індивідуальні значення спостережень, всю автокореляційну структуру часових рядів (для фінансових даних – критично важливу), будь-які залежності, що не вкладаються в перші чотири моменти. Адекватним було б продемонструвати збереження якогось конкретного функціоналу на оригінальних та CDR-представлених даних.

5. Дисертаційна робота добре стилістично і грамотно оформлена, хоча і містить незначну кількість технічних неточностей.

Вважаю, що висловлені зауваження не є визначальними і не применшують загальної наукової новизни і практичної значущості результатів дисертаційного дослідження та не впливають на позитивну оцінку дисертаційної роботи.

Загальні висновки.

Подана до захисту дисертаційна робота є завершеною науковою працею, яка містить нові, добре обґрунтовані результати. Дисертація розв'язує актуальну науково-прикладну задачу обробки даних в Big Data. Тема, зміст та результати дисертації повністю відповідають спеціальності 121 – Інженерія програмного забезпечення.

Зважаючи на актуальність теми, обґрунтованість наукових положень, поданих у дисертації, їхню новизну та практичну цінність, достатню кількість наукових публікацій, а також відсутність порушень академічної доброчесності, вважаю, що дисертаційна робота «Оптимізація структури гетерогенних даних в Big Data» цілком відповідає вимогам пунктів 6 – 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 (зі змінами).

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

Офіційний опонент:
професор кафедри комп'ютерних систем,
мереж та кібербезпеки
Національного університету біоресурсів
і природокористування України,
доктор технічних наук, професор **Олена КРИВОРУЧКО**

Підпис професора Криворучко О. В. засвідчую

Учений секретар _____ ***/Оксана БАРАНОВСЬКА /***

«30» 04 2026р.