РЕЦЕНЗІЯ

доктора технічних наук, доцента,
доцента кафедри комп'ютерних систем та мереж
навчально-наукового інституту фізико-технічних та комп'ютерних наук
Чернівецького національного університету імені Юрія Федьковича
Баловсяка Сергія Васильовича
на дисертаційну роботу аспіранта **Кириченка Євгена Олександровича**
«Оптимізація структури гетерогенних даних в Big Data»,
подану на здобуття наукового ступеня доктора філософії
з галузі знань 12 – «Інформаційні технології»
за спеціальністю 121 – «Інженерія програмного забезпечення»

Актуальність дисертаційного дослідження

Важливою особливістю сучасних інформаційних технологій є потреба в обробці значних обсягів гетерогенних даних, які отримуються з різних джерел, наприклад, із хмарних сховищ. Для таких даних застосування класичних методів аналізу, зокрема класифікації та кластеризації, є ускладненим за рахунок різних типів даних та відсутності їх нормування. З метою підвищення швидкодії аналізу даних при вирішенні прикладних завдань часто виконується обробка не повних копій інформації, а її компактних представлень, які описують ключові статистичні властивості даних. За рахунок обробки компактних представлень даних можливо ефективно аналізувати навіть гетерогенні дані без їх попередньої обробки. Проте, на даний час потребують удосконалення наявні підходи до аналізу слабо структурованих наборів даних та побудови їх компактних представлень, які б враховували різні типи змінних. Зокрема, існує потреба в побудові графових моделей даних та аналізі їх подібності на основі певних метрик. Вирішенню таких задач присвячена дисертаційна робота Кириченко Євгена Олександровича, що зумовлює її актуальність і відповідність до сучасних тенденцій в обробці даних великого обсягу (Big Data).

Зв'язок роботи з державними програмами, планами, темами

Дисертаційне дослідження Кириченко Є.О. виконано на кафедрі програмного забезпечення комп'ютерних систем навчально-наукового інституту фізико-технічних та комп'ютерних наук Чернівецького

національного університету імені Юрія Федьковича за держбюджетною тематикою: «Дослідження, моделювання та розробка програмного забезпечення складних динамічних систем» (Державний реєстраційний номер 0121U109232) та на кафедрі математичних проблем управління і кібернетики Чернівецького національного університету імені Юрія Федьковича за держбюджетною тематикою: «Інформаційні технології в аспекті сучасних задач прийняття рішень» (Державний реєстраційний номер 0121U109159). Дослідження, викладені в дисертаційній роботі Кириченко Євгена Олександровича, відповідають напрямам науково-дослідних робіт кафедри програмного забезпечення комп'ютерних систем та кафедри математичних проблем управління і кібернетики.

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

Наукові положення, висновки та результати дисертації Кириченко Євгена Олександровича обґрунтовані за допомогою загальноприйнятих методів теорії ймовірностей, математичної статистики та машинного навчання, які використано для формування та оцінювання характеристик компактних представлень. Застосовано новітні хмарні технології Amazon Web Services (AWS) для програмної реалізації запропонованої інформаційної технології. Аналіз роботи показує, що автор володіє сучасними методами наукового дослідження та цілеспрямовано їх застосовує. Результати дисертаційної роботи подані чітко, логічно та аргументовано. У дисертації здійснено ґрунтовний огляд літературних джерел, релевантних за тематикою дослідження. Розглянуте дисертаційне дослідження є самостійною науковою працею.

Результати, отримані у під час виконання дисертаційного дослідження, опубліковано у 13 наукових роботах. Основні результати опубліковані у п'яти наукових статтях (2 статті – в журналах, які індексуються у наукометричній базі SCOPUS, 3 статті – в українських фахових виданнях), а також у матеріалах шести міжнародних наукових конференцій. Додатково результати дисертації представлено в одній науковій праці.

Структура дисертації

Дисертація складається зі анотації, вступу, чотирьох розділів та висновків до них, загальних висновків, переліку використаних джерел (232 джерела), чотирьох додатків та списку публікацій автора за темою дисертації (13 наукових робіт). Робота викладена на 224 сторінках. Основні результати дисертації у повній мірі відображені у публікаціях автора.

У вступі обґрунтовано актуальність задачі дослідження, описано мету, завдання, предмет, об'єкт та методи дослідження, висвітлено наукову новизну, теоретичне та практичне значення отриманих результатів, наведено структуру дисертаційної роботи.

У першому розділі дисертації описано теоретичні основи аналізу великих та гетерогенних даних (Big Data), обґрунтовано застосування спектральних методів та апарату випадкових матриць для дослідження даних. Проаналізовано особливості аналізу даних великої розмірності, обґрунтовано потребу в розробці нових методів обробки великих та гетерогенних даних. Розглянуто можливі сфери застосування для методів аналізу великих даних, зокрема, у фінансових і технологічних системах.

Другий розділ присвячений розробці математичного та алгоритмічного апарату для інформаційної технології, яка призначена для компактного подання, класифікації та порівняння великих гетерогенних наборів даних. На основі аналізу сучасних концепцій Big Data запропоновано узагальнену модель компактного представлення даних (Compact Data Representation – CDR), яка дає змогу замінювати повні набори даних їх стислими дескрипторами. Завдяки застосуванню дескрипторів значно зменшується обсяг інформації, необхідний для опису статистичних і структурних характеристик даних. Введено концепцію матриць структурної подібності (Data Information Structure Similarity – DISS) та класи метрик, які дозволяють описувати дані різних типів.

У третьому розділі дисертаційної роботи проведено аналіз сучасних хмарних технологій та масштабованих обчислювальних систем. Розглянуто основні моделі хмарних сервісів (IaaS, PaaS, SaaS) та базові принципи функціонування хмарної інфраструктури. Розроблено інформаційну систему

для обробки Big Data з трирівневою модульною архітектурою. Програмна реалізація інформаційної системи виконана мовою Python із використанням Apache Airflow для оркестрації робочих процесів та Apache Spark для розподіленої обробки даних. Послідовну обробку неструктурованих даних виконано у вигляді програмного конвеєра.

Четвертий розділ присвячений програмній реалізації інформаційної системи та її експериментальній перевірці при обробці гетерогенних табличних даних. Для аналізу даних застосовано метод компактного представлення CDR та метрику структурної подібності DISS. Результати тестування показали, що розроблена інформаційна система може забезпечувати автоматизований аналіз гетерогенних даних на хмарних платформах. За рахунок застосування дескрипторів забезпечується не тільки необхідний рівень точності аналізу даних, але й висока швидкодія їх обробки.

У висновках наведено основні результати дисертаційного дослідження.

Додатки містять список публікацій автора, відомості про апробацію результатів дисертації, акти про впровадження результатів роботи, лістинг частини коду програми.

Наукова новизна

У результаті виконання дисертації отримано результати, які містять такі пункти наукової новизни:

1) вперше:

– розроблено інтегровану модель CDR+DISS, що об'єднує класифікацію змінних, компактне подання даних та структурну метрику подібності для побудови графових моделей корпусу даних;

– запропоновано метод побудови дерева подібностей (similarity tree) на основі мінімального остовного дерева та DISS-метрики, який дозволяє визначати кластерну структуру без попередньої інформації про дані;

– створено багатостадійну хмарно-орієнтовану інформаційну технологію класифікації та кластеризації табличних даних із використанням AWS-сервісів.

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.

Практичне значення одержаних результатів

Практичне значення результатів дослідження полягає у можливості застосування розробленої інформаційної технології для порівняння реальних гетерогенних наборів даних. Застосування методу компактних представлень даних дало змогу зменшити обсяги даних у процесі їх аналізу. У розробленій хмарній архітектурі використано Apache Spark на EMR для розподіленої обробки даних, S3 для зберігання, Airflow для оркестрації, Glue для каталогізації та Athena для ad-hoc запитів. Завдяки застосуванню таких програмних засобів та компактних представлень даних забезпечується розподілена обробка файлів великого розміру з підвищеною на 40-60% швидкодією у порівнянні з традиційними підходами.

Результати дисертаційного дослідження впроваджені в освітній процес Чернівецького національного університету, а також в діяльність двох ТОВ, що підтверджується актами впровадження.

Дискусійні положення та зауваження до змісту дисертаційного дослідження

1. У розділі 2 відсутні схеми для розроблених алгоритмів.

2. У роботі не приведено оцінку часових витрат на обробку даних в розробленій інформаційній технології, проте, такі часові оцінки були б корисними в процесі практичного застосування розроблених програмних засобів.

3. У дисертації проведено аналіз фінансових даних, проте, для демонстрації можливостей розробленої інформаційної технології було б доцільним детальніше висвітлити результати обробки гетерогенних даних інших типів.

4. У таблиці 4.5 для поля «Space Saved» було б доцільно використати більшу кількість цифр після коми з метою кращого розрізнення результатів обробки даних.

Вказані зауваження не применшують наукового значення дисертаційного дослідження Кириченка Євгена Олександровича та не впливають на його загальну позитивну оцінку як самостійного і завершеного.

Загальний висновок

Дисертаційна робота Кириченка Євгена Олександровича є актуальною, має високу теоретичну та практичну цінність. Висновки й основні положення дисертації є обгрунтованими і мають наукову новизну. Під час аналізу матеріалів дисертаційної роботи порушень академічної доброчесності не виявлено. Розглянута дисертаційна робота є завершеним науковим дослідженням, у якому зроблено значний внесок у розвиток методів аналізу гетерогенних табличних наборів даних із використанням хмарної інформаційної технології.

Дисертаційна робота Кириченка Євгена Олександровича «Оптимізація структури гетерогенних даних в Big Data», подана на здобуття наукового ступеня доктора філософії за спеціальністю 121 – «Інженерія програмного забезпечення» в галузі знань 12 – «Інформаційні технології» за її актуальністю, науково-теоретичним рівнем, новизною розв'язання поставлених завдань, практичним значенням отриманих результатів цілком відповідає пунктам 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 постановами Кабінету Міністрів України № 44 від 12.01.2022 р. (зі змінами, внесеними згідно з Постановою Кабінету Міністрів України № 341 від 21.03.2022 р., № 502 від 19.05.2023 р., № 507 від 03.05.2024 р.), а також «Вимогам до оформлення дисертації», затверджених Наказом

Міністерства освіти і науки України №40 від 12.01.2017 р.

Вважаю, що Кириченко Євген Олександрович заслуговує на присудження йому наукового ступеня доктора філософії за спеціальністю 121 – «Інженерія програмного забезпечення» в галузі знань 12 – «Інформаційні технології».

Рецензент
доктор технічних наук, доцент,
доцент кафедри комп'ютерних систем та мереж
навчально-наукового інституту фізико-технічних
та комп'ютерних наук
Чернівецького національного університету
імені Юрія Федьковича Сергій БАЛОВСЯК

Підпис Баловсяка С. засвідчую
Учений секретар Чернівецького національного
університету імені Юрія Федьковича
Ковбовська Н.
04 травня 20