ВІДГУК ОФІЦІЙНОГО ОПОНЕНТА

доктора технічних наук, професора Жолткевича Григорія Миколайовича на дисертацію Коси Вікторії Вікторівни «Метод експериментального дослідження термінологічного насичення в колекціях документів для здобуття знань», що подано на здобуття наукового ступеня доктора філософії зі спеціальності 122 Комп'ютерні науки

Дисертаційну роботу присвячено вирішенню важливої та раніше не вирішеної науково-практичної задачі – розробці ефективного та результативного комплексного обчислювального методу для експериментального дослідження колекцій професійних документів у межах довільної предметної області з метою виявлення та вимірювання термінологічного насичення.

Ступінь актуальності обраної теми

Загальним контекстом когнітивної діяльності сучасної людини є протиріччя між обмеженістю швидкості сприйняття інформації (в середньому 25 біт за секунду) і неперервно зростаючим обсягом інформації, яку треба приймати до уваги в процесі прийняття рішень. Це загальне протиріччя є джерелом багатьох негативних ефектів, що спостерігаються у різноманітних сферах діяльності.

В ситуації опанування нової предметної області, яка є типовою для освіти, наукової діяльності, інженерії, особливо у сфері інформаційних технологій, дуже важливим є етап об'єктивації і засвоєння термінологічної бази предметної області, і у зв'язку з цим — оцінка достатності сформованої термінологічної бази для успішної діяльності у контексті предметної області, пов'язаної з побудовою моделей предметної області різного рівня абстракції і комунікування з іншими фахівцями з цієї предметної області.

В контексті дисертаційної роботи, що рецензується, саме ця термінологічна база стає відправним пунктом для побудови онтологій предметної області, які використовуються для подальшої автоматизації процесів систематизації інформації.

Авторка дослідження виходить з гіпотези про те, що достатність термінологічної бази предметної області визначається її "насиченістю", яка формально вимірюється введеною в дисертації характеристикою "термінологічна насиченість". Фактично робота присвячена аналізу поведінки цієї характеристики і експериментальній демонстрації, що гіпотеза дослідження не суперечить відомим фактам.

Таким чином, можна зробити висновок про те, що обрана здобувачкою тема роботи є актуальною і такою, що має суттєву значущість для вирішення важливої науково-технічної задачі у галузі комп'ютерних наук і є абсолютно релевантною спеціальності 122 Комп'ютерні науки.

Структура роботи

Дисертаційна робота складається зі вступу, п’яти розділів, висновків, списку використаних джерел та додатків. Загальний обсяг дисертації складає 242 сторінок, з них: 14 алгоритмів; 43 рисунки; 22 таблиці; список використаних джерел зі 124 найменувань на 15 сторінках; 6 додатків на 32 сторінках. Основна частина дисертації викладена на 167 сторінках (приблизно 6.8 авторських аркушів).

Логічна структура викладення результатів дисертаційного дослідження є цілком традиційною: вступ, розділи основного змісту і висновки.

Вступ до дисертаційної роботи детально обґрунтовує актуальність теми дисертаційної роботи, визначає зв’язок роботи з науково-дослідними темами і науково-технічними проектами, формулює мету і завдання дослідження, визначає об’єкт, предмет та методи дослідження, формулює наукову новизну та практичне значення отриманих результатів, містить інформацію про практичне використання результатів роботи, визначає особистий внесок здобувача, містить інформацію про апробацію результатів дослідження та їх висвітлення у наукових публікаціях.

Розділ 1 присвячений аналізу сучасного стану досліджень за тематикою роботи. Він розпочинається з опису методики, що використовується в роботі для відбору для детального огляду і аналізу літературних джерел. Далі ця методика застосовується для аналізу сучасного стану досліджень релевантних предметних областей. Цей аналіз дозволив авторці виявити відкриті питання та рівень важливості вирішення цих питань, а також сформулювати мету і основні завдання дисертаційного дослідження.

Розділ 2 присвячений першому завданню дослідження – розробці формального фреймворку для методу виявлення та вимірювання термінологічного насичення. Розділ починається з викладення базових визначень та відомих фактів. На цій основі сформульовано гіпотезу дослідження, несуперечливість якої підтверджується теоретичним дослідженням, що містить

– формальне визначення функції термінологічної різниці і доведення її метричних властивостей у просторі усіх можливих колекцій документів для домену;

– формулювання та доведення теореми про достатні умови існування термінологічного насичення;

– викладення формального підходу щодо оптимізації методу вимірювання та виявлення термінологічного насичення, доведення коректності розробленого оптимізованого методу, обґрунтування можливостей паралельної реалізації.

Розділ 3 присвячений другому завданню дослідження – розробці нових або вдосконаленню раніше розроблених алгоритмів, що імплементують розроблений формальний фреймворк (поданий у розділі 2) у вигляді обчислювального методу. Починає розділ викладення концепції розробленого конвеєру обробки колекцій документів. Далі наводяться розроблені або вдосконалені в роботі алгоритми, що забезпечують функціональність модулів конвеєру. Завершується розділ описом програмної реалізації розроблених

алгоритмів. В тексті є діюче посилання на відповідне програмне забезпечення, що дозволяє як перевіряти результати авторки, так і використовувати її ідеї в інших дослідженнях і освітньому процесі.

Розділ 4 присвячений третьому завданню дослідження – експериментальній оцінці розробленого комплексного методу, імплементованого набором алгоритмів. У розділі сформульовано завдання експериментального дослідження та план проведення експериментів, після чого наведено їх результати. Завершується розділ аналізом результатів проведених експериментів, який демонструє коректність, незалежність від домену, ефективність та масштабованість розробленого обчислювального методу, набору алгоритмів, що його реалізує, та відповідного програмного забезпечення.

Розділі 5 присвячений четвертому завданню дослідження – практичному використанню науково технічних результатів дослідження у промисловості та академічній практиці. Набутий досвід узагальнюється шляхом висвітлення потенційних переваг розробленого методу та програмного забезпечення для промислових користувачів, представлення потенційних бізнес-сценаріїв щодо застосування результатів роботи у галузі наукового видавництва.

Висновки по роботі узагальнюють та фокусують увагу на основних отриманих результатах, а також містять плани щодо подальшого розвитку науково-технічного доробку авторки.

Дисертація є структурно та змістовно збалансованою. Послідовність викладення її положень є логічною і такою, що побудована у відповідності з науковим методом.

Ступінь обґрунтованості наукових положень, висновків і рекомендацій та їх достовірність

Ступінь обґрунтованості та достовірності результатів дисертаційного дослідження не викликає сумнівів, що забезпечується

– обраною методологією дослідження;

– аналізом стану досліджень в предметній області дослідження, комплексним застосуванням теоретичних та експериментальних методів дослідження;

– формальним, а в разі його складності експериментальним, доведенням усіх положень та рекомендацій;

– можливістю відтворення результатів роботи за рахунок надання публічного доступу до усіх необхідних компонентів, включаючи дані та інструментальне програмне забезпечення;

– адекватною добіркою використаних літературних джерел;

– успішними упровадженнями науково-технічних результатів дослідження у промисловості та вищій школі (підтверджено двома довідками про впровадження).

Наукова новизна отриманих результатів

У дисертаційному дослідженні отримано наступні нові науково-технічні результати.

Уперше розроблено комплексний обчислювальний метод виявлення та вимірювання термінологічного насичення у послідовності інкрементально зростаючих підколекцій гіпотетично існуючої повної колекції професійних документів, що описують довільний домен. (Розроблений метод є різновидом методу послідовного наближення у метричному просторі колекцій документів, що є підмножинами повної колекції; доведено теорему існування термінологічного насичення, що надає достатні умови збіжності методу; розроблено алгоритми та програмне забезпечення для імплементації цього обчислювального методу).

Отримали подальший розвиток формальне визначення міри термінологічної різниці між двома наборами термінів зі значеннями оцінок їх значущості (формально доведено її метричні властивості; обчислювальний метод автоматичного здобуття термінів на базі методу C-value – запропоновано, замість обчислення C-value термінів, здобутих з усієї підколекції документів, обчислювати часткові C-value, здобуті з інкременту колекції документів, і потім зливати часткові C-value; доведено, що злиті часткові C-value практично не відрізняються від C-value, обчислених попереднім методом; експериментально доведено, що розвинутий метод є коректним, ефективним, таким що очевидним чином розпаралелюється, та масштабованим для обробки текстів будь-якого великого обсягу).

Удосконалено обчислювальний конвеєр виявлення, вимірювання та аналізу термінологічного насичення, що використовує розроблений метод виявлення та вимірювання термінологічного насичення та розвинутий метод автоматичного здобуття термінів з використанням злитих часткових C-value шляхом: залучення обчислювального методу для відбору релевантних документів до інкрементів колекції; використання розробленої техніки та алгоритмів групування частково подібних термінів; впорядкування документів для формування інкрементів колекції за зменшенням частоти цитування документів.

Отримані нові результати, у сукупності, розв’язують важливу науково-технічну задачу виявлення та вимірювання термінологічного насичення у текстових колекціях для оцінки ступеню репрезентативності текстових колекцій для здобуття знань для обраній предметній області.

Значимість результатів дослідження для науки і практики та можливі шляхи їх використання

Наукова та практична значимість результатів дисертаційного дослідження підтверджується тим, що розроблений комплексний метод є повністю реалізованим у програмному забезпеченні, а авторкою надано усі необхідні умови задля наукового відтворення цих результатів іншими вченими, тобто можливість використання розробленого методу іншими науковцями, у їх

власних дослідженнях і можливість оцінювання розробки у порівнянні з розробками інших авторів.

Висока значимість для науки отриманих у роботі результатів підтверджується рівнем цитувань публікацій здобувача (див. додаток Ж дисертації).

Практичне значимість результатів дисертаційної роботи підтверджується впровадженням у промисловому та академічному контекстах (відповідні довідки наведені у Додатку Д до дисертації.

Повнота викладу результатів дослідження в наукових публікаціях

Основні наукові результати роботи в повному обсязі викладені у 7 публікаціях (у співавторстві). Інформація про особистий внесок здобувача у ці публікації наведено у дисертації, що дозволяє зробити висновок про його достатність.

З цих робіт, 7 статей опубліковано у міжнародних періодичних виданнях, що проіндексовані у наукометричній базі «Scopus».

Наукові результати, що отримані у дисертаційній роботі, апробовані та отримали позитивну оцінку на 2 міжнародних конференціях, 2 міжнародних симпозіумах та конкурсі молодих науковців Запорізької обласної державної адміністрації.

Академічна доброчесність

Проведений аналіз тексту (в тому числі з використанням інструментів, що рекомендовані для використання у Харківському національному університеті імені В.Н. Каразіна) дає підстави для висновку про відсутність порушень академічної доброчесності.

Відповідність дисертації вимогам, передбаченим пунктом 10 Порядку проведення експерименту з присудження ступеня доктора філософії

Дисертацію подано у вигляді спеціально підготовленої кваліфікаційної наукової праці на правах рукопису, що виконувалася здобувачем особисто. Дисертація містить наукові положення, нові науково обґрунтовані теоретичні та експериментальні результати проведених здобувачем досліджень, що мають істотне значення для галузі комп'ютерних наук. Це підтверджено публікаціями, що розкривають основний зміст роботи. Дисертація свідчить про суттєвий особистий внесок здобувача в науку та характеризується єдністю змісту.

Дисертацію оформлено у відповідності до вимог Міністерства освіти і науки України (Наказ №40 від 12.01.2017 із змінами, внесеними згідно з Наказом Міністерства освіти і науки № 759 від 31.05.2019).

Дискусійні положення та зауваження до змісту дисертації

Наведені вище аргументи на користь позитивної оцінки дисертаційної роботи не виключають зауважень і відкритих питань, зокрема:

1. У розділі 2 дисертації доведено теорему про існування термінологічного насичення в колекції документів. Ця теорема надає достатні умови існування термінологічного насичення. Однак, як відзначає авторка, цей важливий резуль-

тат не дозволяє прогнозувати досягнення термінологічного насичення за даними декількох перших вимірювань термінологічної різниці. Було б доцільним розробити метод такого прогнозу, що суттєво підвищило б практичну значущість результату роботи.

2. Задача розробки методу ставиться у банаховому просторі, який авторка некоректно називає гільбертовим. У доведенні специфічні властивості гільбертових просторів не використовуються, тому вважаю, що використання поняття гільбертового простору є недоречним.

3. Обрання у якості базової міри для вимірювання термінологічної різниці між наборами термінів Манхеттен відстані здається таким, що не відповідає найбільш поширеним підходам у галузі. Більшість наукових та промислових розробок використовують косинусну міру. Таким чином, вибір іншої мірі погіршує можливість співставлення результату роботи з іншими розробками.

4. Доведення теореми про еквівалентність звичайних C-value та злитих частковим C-value виглядає нетрадиційно. Дійсно, твердження теореми є доведеним, якщо вірною є гіпотеза H_1. Однак, доведення H_1 виконується не формально, що було б очікуваним, а експериментально. Було б доцільно дослідити вірність гіпотези H_1 формально, а потім перевірити, чи відповідають результати експерименту формально доведеним положенням.

5. Метою промислового проекту, через який було впроваджено науково-технічний доробок дисертації, було відповісти на три питання (додаток Е). У розділі 5.1 дані відповіді на перши два з цих питань, а трете залишилося без відповіді. Чи розроблений у дисертації метод є спроможним надати відповідь на трете питання цього промислового проекту?

6. В опису академічного впровадження науково-технічного доробку дисертації вказано, що тільки чотири з дев'ятнадцяти студентів використали конвеєр з вимірювання термінологічного насичення для перевірки репрезентативності відібраної літератури. По-перше, загальна кількість студентів замала для того, щоб судження про результати застосування було статистично достовірним. По-друге, тільки 12.1% студентів з тих, хто брав участь, використали конвеєр вимірювання термінологічного насичення. Чи можна за цими показниками вважати, що метод був добре сприйнятий студентами?

7. В огляді літературних джерел зазначено, що феномен термінологічного насичення при автоматичному здобутті термінів з колекцій документів досі не досліджувався. Це виглядає дуже сильним твердженням. Чи є доказ того, що при аналізі літератури якась робота не була пропущена? З іншого боку, якщо ніхто не вивчав цього феномену, це дослідження не було нікому цікавим?

8. На мою думку формулювання наукової новизни можна було б виконати у більш традиційний спосіб, що б сфокусувало увагу саме на результаті, а не на його мотивації.

9. Робота містить друкарські помилки, проте їх кількість не перевищує встановлені стандартом норми, а також в роботі де-інде використовуються стилістично невдалі конструкції.

Загальний висновок

Зазначені зауваження не є такими, що суттєво впливають на загальну позитивну оцінку представленого наукового дослідження. Вважаю, що дисертація на тему «Метод експериментального дослідження термінологічного насичення в колекціях документів для здобуття знань» є завершеною самостійною науковою роботою, яка містить нові аргументовані результати в галузі комп'ютерних наук і за актуальністю, змістом, науковою новизною, обґрунтованістю висновків, достовірністю і значущістю відповідає вимогам «Порядку проведення експерименту з присудження ступеня доктора філософії», затвердженому Постановою Кабінету Міністрів України № 167 від 6 березня 2019 року, а її авторка, Коса Вікторія Вікторівна, заслуговує на присудження наукового ступеню доктора філософії за спеціальністю 122 Комп'ютерні науки.

Офіційний опонент:
доктор технічних наук, професор, декан
факультету математики і інформатики,
професор кафедри теоретичної та
прикладної інформатики Харківського
національного університету імені В.Н. Каразіна Григорій ЖОЛТКЕВИЧ
"___" _____ 2021 року

Підпис професора Жолткевича Григорія Миколайовича засвідчую
Начальник служби управління персоналом
Харківського національного університету
імені В.Н. Каразіна, професор Сергій КУЛІШ
"___" _____ 2021 року