ВІДГУК

офіційного опонента, доктора технічних наук, професора
Яковини Віталія Степановича на дисертацію Коси Вікторії Вікторівни
«Метод експериментального дослідження термінологічного насичення
в колекціях документів для здобуття знань», поданої на здобуття наукового
ступеня доктора філософії зі спеціальності 122 – «Комп'ютерні науки»

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

Ступінь актуальності обраної теми

Актуальність обраної теми зумовлена тим, що отримання набору термінів з колекцій документів у межах окремої предметної області є початковим етапом здобуття знань для побудови онтології обраної предметної області. Онтології предметних областей є формальними описовими теоріями для подання знань у різноманітних застосуваннях та розробках засобів штучного інтелекту, що сьогодні є одним з пріоритетних напрямків розвитку світової та вітчизняної науки. Актуальною проблемою у побудові онтологій шляхом їх вивчення з текстів, яка стримує практичне їх застосування, є невизначеність того, як забезпечувати репрезентативність вихідних колекцій документів. У даній дисертації зроблено суттєвий крок на шляху вирішення цієї проблеми – розроблено комплексний обчислювальний метод, що дозволяє виявляти термінологічно насичені підколекції документів шляхом експериментального дослідження термінологічного насичення у текстах вихідної колекції.

Таким чином, обрана здобувачем тема роботи є актуальною і такою, що має суттєву значущість для вирішення сучасних проблем штучного інтелекту та його практичних застосувань. Обрана тема відповідає спеціальності 122 – «Комп'ютерні науки» (з урахуванням відсутності затвердженого стандарту третього (освітньо-наукового) рівня вищої освіти).

Структура роботи

Дисертаційна робота складається зі вступу, п'яти розділів, висновків, списку використаних джерел та додатків. Загальний обсяг дисертації складає 244 сторінки, з них: 14 алгоритмів; 43 рисунки; 22 таблиці; список використаних джерел зі 124 найменувань на 15 сторінках; 7 додатків на 35 сторінках. Основна частина дисертації викладена на 167 сторінках.

У вступі обґрунтовано актуальність теми дисертаційної роботи, зазначено зв'язок роботи з науково-технічними проектами, сформульовано мету і завдання

дослідження, визначено об’єкт, предмет та методи дослідження, показано наукову новизну та практичне значення отриманих результатів, наведено інформацію про практичне використання доробку, особистий внесок здобувача, апробацію результатів дослідження та їх висвітлення у наукових публікаціях.

У розділі 1 розглянуто та проаналізовано сучасний стан досліджень за тематикою роботи. Розділ починається з викладу методології, яку було використано для відбору літературних джерел для систематичного огляду і аналізу. Далі розглянуто та проаналізовано сучасний стан досліджень за напрямком дисертаційної роботи. На базі проведеного аналізу представлено виявлені прогалини у сучасному стані досліджень та мотивацію щодо зменшення цих прогалин, що призвело до формулювання питань та постановки завдань дисертаційного дослідження. Завдання сформульовано у відповідності до наукового методу та базуючись на баченні підходу для вирішення поставлених питань дослідження. Особливо хотілось би відзначити Табл. 1.5, в якій підсумовано завдання та питання дослідження, разом з посиланнями на відповіді у відповідних розділах дисертаційної роботи.

У розділі 2 наведено хід та результати виконання першого завдання дослідження – розробки формального фреймворку для методу виявлення та вимірювання термінологічного насичення. Розділ починається з базових відомостей. Далі представлено гіпотези дослідження, які необхідно перевірити в контексті теоретичного підходу. Наступні підрозділи присвячені формальному введенню функції термінологічної різниці (thd) та доведенню її метричних властивостей у просторі усіх можливих колекцій документів для домену; формулюванню та доведенню теореми про достатні умови існування термінологічного насичення; розробці формального підходу щодо оптимізації методу вимірювання та виявлення термінологічного насичення, доведенню коректності розробленого оптимізованого методу, обґрунтуванню його розпаралелюваності та результативності.

У розділі 3 виконується друге завдання дослідження – розробка нових та вдосконалення існуючих алгоритмів, що реалізують розроблений формальний фреймворк (розділ 2) у вигляді обчислювального методу. Розділ починається з викладення розробленого конвеєру обробки колекцій документів. Далі наведені розроблені або вдосконалені алгоритми, що є функціональним змістом модулів конвеєру обробки. Наприкінці розділу представлено програмну реалізацію розроблених алгоритмів та наведено посилання на це програмне забезпечення, що є загальнодоступним для використання у академічних дослідженнях.

Розділ 4 присвячено виконанню третього завдання дослідження – експериментальній оцінці розробленого комплексного методу, реалізованого як набір алгоритмів. На початку розділу сформульовані завдання експериментального дослідження та детальний план проведення експериментів. Далі наведено та проаналізовано результати цих експериментів. Як ґрунтовно пояснюється у розділі, результати доводять правильність, незалежність від домену, ефективність та масштабованість розробленого обчислювального методу, набору алгоритмів та програмного забезпечення.

У розділі 5 представлено виконання четвертого завдання роботи – апробація та узагальнення досвіду використання науково-технічних результатів дослідження у промисловості та академічній практиці. Узагальнення досвіду практичного використання результатів роботи зроблено шляхом: висвітлення потенційних переваг розробленого методу та програмного забезпечення для промислових користувачів; представлення потенційних бізнес-сценаріїв щодо застосування результатів роботи у галузі наукового видавництва.

У висновках наведені загальні висновки до роботи та плани щодо подальшого розвитку науково-технічного доробку дисертації.

Дисертація є структурно та змістовно збалансованою. Послідовність викладення її положень є чіткою, структуровано, 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 є високим. Це забезпечується: адекватно обраною методологією дослідження; систематичним аналізом стану досліджень в обраному напрямку, комплексним застосуванням теоретичних та експериментальних методів дослідження; формальним та експериментальним доведенням усіх положень та рекомендацій; репрезентативною базою використаних літературних джерел; забезпеченою можливістю відтворення результатів роботи за рахунок надання публічного доступу до усіх необхідних компонентів, включаючи дані та інструментальне програмне забезпечення; успішними упровадженнями науково-технічного доробку дисертації у промисловості та вищій школі, про що свідчать дві довідки про впровадження.

Наукова новизна отриманих результатів

У дисертації наведено такі 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. Доведено теорему існування термінологічного насичення, що надає достатні умови існування.

Отримали подальший розвиток:

- Формальне визначення міри термінологічної різниці між двома наборами термінів зі значеннями оцінок їх значущості, зокрема формально доведено її метричні властивості.
- Обчислювальний метод автоматичного здобуття термінів на базі методу C-value, який, на відміну від існуючого, замість обчислення C-value термінів, здобутих з усієї підколекції документів, обчислює часткові C-value, здобуті

з інкременту колекції документів, і потім зливає отримані часткові C-value. Доведено, що злиті часткові C-value практично не відрізняються від C-value, обчислених існуючим методом. Експериментально доведено, що розвинутий метод є коректним, ефективним 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 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 обраної предметної області.

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

Наукова та практична значущість результатів дисертаційного дослідження є високою.

Висока значущість отриманих результатів для науки полягає у формальному математичному доведенні достатніх умов існування термінологічного насичення, метричних властивостей функції термінологічної різниці, індивідуального порогу значущості терміну, рівності *cv* та *mpcv*. Практична цінність витікає з того, що розроблений комплексний метод є повністю реалізованим у програмному забезпеченні та автором забезпечені усі необхідні умови задля наукового відтворення цих результатів іншими вченими. Таким чином, забезпечено можливість: (і) використання розробленого методу іншими науковцями у їх власних дослідженнях; (іі) порівняльної оцінки з розробками інших вчених. Висока значущість для науки отриманих у роботі результатів також підтверджується значною кількістю цитувань публікацій здобувача (додаток Ж дисертації).

Практичне значення результатів дисертаційної роботи підтверджується їх практичним впровадженням у промисловому та академічному контекстах. Впровадження підтверджені відповідними довідками (Додаток Д). Крім того, робота чітко визначає практичні переваги впровадження свого доробку у промисловості та надає потенційні сценарії використання в галузі наукового видавництва.

Повнота викладу результатів дослідження в наукових публікаціях

Основні наукові результати роботи в повному обсязі викладені у семи публікаціях (у співавторстві), що відповідає п. 11 Порядку проведення експерименту з присудження ступеня доктора філософії. Особистий внесок здобувача у ці публікації є суттєвим, що наведено у дисертації. Серед цих робіт, сім 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 двох міжнародних конференціях, двох міжнародних симпозіумах та конкурсі молодих науковців Запорізької обласної державної адміністрації.

Академічна доброчесність

Очевидних ознак порушення автором академічної доброчесності, зокрема випадків оприлюднення, частково або повністю, наукових результатів, отриманих іншими особами, як результатів власного дослідження та/або відтворення опублікованих текстів інших авторів без зазначення їх авторства, не виявлено.

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 ґрунтовно підтверджено публікаціями, що розкривають основний зміст роботи. Дисертація свідчить про суттєвий особистий внесок здобувача в науку та характеризується єдністю змісту. Все вищезазначене дає змогу зробити висновок про відповідність роботи вимогам п. 10 Порядку проведення експерименту з присудження ступеня доктора філософії.

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

Загалом, позитивно оцінюючи наукове і практичне значення отриманих дисертантом результатів, варто відзначити наступні дискусійні положення і зауваження до змісту дисертаційної роботи.

1. Не викликає сумніву твердження щодо того, що підколекція з термінологічним насиченням є репрезентативною щодо опису домену. Однак, на мою думку, в роботі не доведено необхідність існування такої

підколекції, тобто явища термінологічного насичення. І дійсно, як відзначає авторка, достатні умови існування термінологічного насичення не дають змоги прогнозувати досягнення термінологічного насичення за даними декількох перших вимірювань термінологічної різниці. Було б доцільним дослідити необхідність існування явища термінологічного насичення та розробити метод прогнозу його досягнення, що суттєво підвищило б практичну значущість результатів роботи.

2. В роботі пропонується обчислювати значення *score* простою більшістю голосів. На мою думку, в такому випадку важливі, проте вузькоспеціалізовані терміни, які не часто вживаються, можуть випасти з колекції.
3. На стор. 76 роботи пропонується обирати значення *eps* методом простого голосування. Однак в такому випадку 50% - 1 термінів не враховуються в наборі збережених значущих термінів. Чи коректним є такий підхід? Чи автор розглядала можливість використання, наприклад, якихось статистичних критеріїв, пов'язаних з розподілом чи частотою вживання термінів, наприклад 3σ?
4. В теоремі 2.3, стор. 88, на мою думку, третя умова сформульована не зовсім коректно. З формулювання можна припустити, що функція $eps_{min}(i)$ на усій області визначення не менша за $thds_{max}(i)$, в той час як, з урахуванням першої і другої умов, достатньо, щоб хоча б в одній точці виконувалась ця умова, тобто умову можна записати як $\exists\, j\colon eps_{min}(j) \geq thds_{max}(j)$, власне як це і написано на стор. 89.
5. На стор. 90 роботи наведений виграш в часі виконання є не у $\left(\frac{n+1}{2} - 1\right) \cdot k$ разів, а на зазначену величину – в роботі наведено різницю, а не частку.
6. З підрозділу 3.6.2 та таблиць 3.4 і 3.5 не зрозуміло наскільки універсальними є визначені пороги подібності термінів: чи це застосовно тільки для дослідженого тестового набору чи може бути використано для будь-якої множини термінів і текстів.
7. Стиль та оформлення роботи, на мою думку, може бути покращений шляхом використання безособових дієслівних форм та речень, більш характерних для наукового стилю; вживання застандартованої наукової та/чи науково-технічної термінології, запровадженої національними стандартами на терміни та визначення понять; згідно ДСТУ 3008:2015 розділ «Висновки» не слід нумерувати; виправленням граматичних помилок та описок, які зустрічаються в тексті, наприклад замість «строка» слід вживати «рядок», замість «кінцевим» на стор. 73 – «скінченним» тощо.

Загальний висновок

Зазначені зауваження та рекомендації не 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 обґрунтовані теоретичні та експериментальні

результати, що мають істотне значення для 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 автор, Коса Вікторія Вікторівна, заслуговує, за результатами публічного захисту наукових досягнень у формі дисертації, на присудження наукового ступеня доктора філософії за спеціальністю 122 – «Комп'ютерні науки».

ОФІЦІЙНИЙ ОПОНЕНТ:

Доктор технічних наук, професор
професор кафедри систем штучного
інтелекту
Інституту комп'ютерних наук та
інформаційних технологій
Національного університету
«Львівська політехніка» Яковина В.С.

Підпис Яковини В.С. підтверджую

Вчений секретар
Національного Університету
«Львівська Політехніка» Брилинський Р.Б.