

ВИСНОВОК

ПРО НАУКОВУ НОВИЗНУ, ТЕОРЕТИЧНЕ ТА ПРАКТИЧНЕ ЗНАЧЕННЯ РЕЗУЛЬТАТІВ ДИСЕРТАЦІЇ

Коси Вікторії Вікторівни «Метод експериментального дослідження термінологічного насичення в колекціях документів для здобуття знань», що подана на здобуття наукового ступеня доктора філософії за спеціальністю 122 – «Комп'ютерні науки» (галузь знань 12 – «Інформаційні технології»)

Дисертація Коси Вікторії Вікторівни «Метод експериментального дослідження термінологічного насичення в колекціях документів для здобуття знань», що подана на здобуття наукового ступеня доктора філософії за спеціальністю 122 – «Комп'ютерні науки» (галузь знань 12 – «Інформаційні технології») виконана на кафедрі комп'ютерних наук Запорізького національного університету Міністерства освіти і науки України. Тема дисертації затверджена за засіданні Науково-технічної ради Запорізького національного університету (протокол № 4 від 29 листопада 2016 року)

Для підготовки висновку про наукову новизну, теоретичне та практичне значення результатів дисертації «Метод експериментального дослідження термінологічного насичення в колекціях документів для здобуття знань» Вченою радою Запорізького національного університету (22 грудня 2020 року) визначено, що попередня експертиза дисертації проводитиметься на базі математичного факультету Запорізького національного університету, та призначено двох рецензентів:

1) професора кафедри економічної кібернетики Запорізького національного університету, доктора фізико-математичних наук, професора **Козіна Ігоря Вікторовича**;

2) доцента кафедри програмної інженерії Запорізького національного університету, доктора технічних наук, доцента **Чопорова Сергія Вікторовича**.

1. Ступінь актуальності теми дослідження

Дисертаційну роботу присвячено вирішенню комплексної 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.

Колекції професійних документів, що описують ту чи іншу предметну область, все частіше використовуються для здобуття з них знань для побудови описових теорій (онтологій) для предметної області. Невирішена проблема полягає у відсутності об'єктивного методу забезпечення репрезентативності вихідної колекції документів. Розроблений у дисертації комплексний обчислювальний метод базується на тому, що семантичним відбитком документу, у деякій

професійній предметній області, є набір використаних документом термінів. Тому, семантичним відбитком колекції документів є набір термінів, здобутих з усіх її документів. Виходячи з цього, колекція буде репрезентативно описувати предметну область тоді, коли вона містить підколекцію з практично тим самим набором термінів, що і уся колекція. Така підколекція у роботі визначається такою, у якій спостерігається термінологічне насичення. Виявлення та вимірювання цього термінологічного насичення актуально і важливо тому, що: (i) дає змогу відокремити репрезентативну підколекцію мінімального обсягу, якщо вона існує, що гарантує якісний результат у подальшому процесі вивчення онтології з текстів; та (ii) підвищує ефективність здобуття термінів та знань, тому, що реальні колекції документів для професійних предметних областей є дуже великими.

2. Об'єкт, предмет, мета та завдання роботи

Об'єктом дослідження є процес автоматизованого здобуття, з колекцій релевантних документів, наборів термінів, що характеризують довільну професійну предметну область, для подальшої побудови онтологій цієї предметної області, з урахуванням впливу явища термінологічного насичення.

Предметом дослідження є метод експериментального дослідження термінологічного насичення в колекціях документів для здобуття знань у процесі вивчення онтологій довільного домену.

Метою роботи є підвищення репрезентативності, ефективності та результативності здобуття термінології з колекцій професійних документів у межах довільної предметної області для подальшої побудови онтологій, шляхом розробки комплексного обчислювального методу виявлення та вимірювання термінологічного насичення в колекціях професійних текстових документів, що описують предметну область.

Для досягнення мети в роботі, на основі систематичного огляду та аналізу літературних джерел, що розкривають сучасний стан досліджень в обраній області, поставлені та вирішені наступні **завдання**:

1) розробити формальний фреймворк методу виявлення та вимірювання термінологічного насичення, що охоплює визначення, формалізми та докази всіх його компонентів, необхідних для побудови якісного, ефективного та результативного обчислювального методу;

2) спираючись на розроблений формальний фреймворк, розробити алгоритми що матеріалізують **обчислювальний метод** для здобуття насиченої термінології; реалізувати ці алгоритми у програмному забезпеченні та надати програмне забезпечення для академічного та промислового використання;

3) використовуючи розроблене програмне забезпечення, експериментально оцінити та перевірити розроблений обчислювальний метод виявлення та вимірювання термінологічного насичення на синтетичних та реальних колекціях наукових публікацій, що належать до різних доменів.

4) проаналізувати, якими є потенційні практичні наслідки (переваги) використання розробленого методу на основі досвіду реалізації кейсів

практичного використання в промисловості та академічній сфері; проаналізувати потенційні переваги впровадження науково-технічного доробку та потенційні бізнес-сценарії у цільовій промисловій галузі наукового видавництва.

3. Методи дослідження

Досягнення мети та виконання завдань дисертаційного дослідження здійснено за допомогою комплексного використання: наукового методу; формальних математичних методів; методів та формальних мов специфікації обчислювальних процесів та алгоритмізації; методів та мов програмування; методів обробки текстових даних; методів та мір аналізу текстових даних і термінів; методів планування, виконання та аналізу обчислювальних експериментів.

4. Рівень обізнаності здобувача про сучасний стан досліджень у контексті роботи

Науковим підґрунтям для виконання дисертаційної роботи слугували наукові праці українських та закордонних фахівців, що були систематично досліджені та проаналізовані. Досліджувався сучасний світовий науковий доробок у наступних галузях знань та наступних авторів, що достатньо покриває існуючий стан наукових досліджень у контексті дисертації:

- розробка онтологій домену та здобуття вимог (A. Gómez-Pérez, N. Guarino, S. Pinto, C. Sarasua, G. Schreiber, R. Studer, M. C. Suárez-Figueroa, Y. Sure, B. Єрмолаєв, О. Татаринцева);
- Вивчення онтологій з текстів та консенсус спільноти (P. Buitelaar, G. Corpas Pastor, A. Maedche, D. Maynard, E. Motta, F. Osborne, A. Salatino, M. Seghiri Domínguez, S. Staab, W. Wong);
- збір релевантних документів доброї якості (A. Ahad, K. Beatty, J. Lescy, A. Medlar, A. Varela, H. Waddington, Г. Добровольський, Н. Кеберле);
- термінологічне насичення та репрезентативність (A. Berenstein, L. Chernyak, J. Doerre, A. Ferrari, H. Han, A. Voronkov);
- теоретичне насичення та вивчення онтологій (C. Aldiabat, B. Glaser, C.-L. Le Navenec, A. Strauss, О. Татаринцева);
- впорядкування документів для обробки (J. Doerre, H. Pohl, A. Mottelson, G. J. Schneider, R. Costas, W. Glänzel, A. Schubert, P. Savov);
- автоматизовані методи здобуття термінів (K. Ahmad, S. Ananiadou, N. Astrakhantsev, G. Bordea, C. Badenes-Olmedo, K. Church, O. Corcho, B. Daille, T. Dunning, D. Evans, I. Fahmi, K. Frantzi, W. Gale, U. Hahn, P. Hanks, I. Korkontzelos, L. Kozakov, R. Lefferts, J. Lossio-Ventura, N. Loukachevitch, D. Manning, D. Maynard, O. Medelyan, M. Nokel, Oliver, Park, Peñas, Schutze, Sclano, V`azquez, Velardi, J. Wermter, I. Witten, Z. Zhang);
- вимірювання подібності текстів (M. Arnold, L. Dice, C. Elkan, A. Fahmy, W. Goma, R. Hamming, A. Huang, P. Jaccard, M. Jaro, H. Lee, V. Levenshtein, C. Lu, G. Miller, A. Monger, E. Ohlebusch, J. Qin, A. Singhal, T. Sørensen, Y. Tsuruoka, W. Winkler, M. Yu);
- ефективне співставлення строк для пошуку вкладених термінів (A. Aho, A.

Browne, F. Chowdhury, M. Corasick, R. Farrell, C. Lu, Z. Zang).

У роботі було використано автоматизований підхід до пошуку та забезпеченню повноти вибору релевантних публікацій для аналізу літературних джерел, що обґрунтовує систематичність та повноту обізнаності автора з сучасним станом науково-технічного доробку в контексті дисертаційного дослідження.

Водночас, враховуючи все розмаїття наукових праць з окремих питань проблематики здобуття термінів та вимог до побудови описових теорій предметних областей, й жодним чином не применшуючи їх значення, варто зазначити, що комплексних наукових досліджень, безпосередньо присвячених методам здобуття репрезентативних наборів термінів для подальшого здобуття знань та побудови онтологій, у контексті існування, виявлення та вимірювання термінологічного насичення у професійних колекціях документів не було проведено. Тому, дисертаційна робота є новаторською у цьому науково-технічному напрямку, вищезазначені факти актуалізують представлене в дисертації дослідження та зумовлюють його теоретичне та практичне значення.

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

Здобувач вірно визначає зв'язок роботи з науковими програмами, планами, темами, грантами, вказує на те, що дослідження проводились у рамках виконання науково-дослідної тематики кафедри комп'ютерних наук Запорізького національного університету та кооперації з промисловістю регіону:

- Міжнародному проєкту EU FP7 Marie Curie IRSES SemData project (<http://www.semdata-project.eu/>), grant agreement No PIRSES-GA-2013-612551. Строки виконання проєкту: 10.2013 – 10.2017. Науковий керівник від України: к.ф.-м.н., доц. Єрмолаєв В. А. Участь здобувача – виконавець.

- Промисловому проєкту Strategic Analysis of R&D Gaps and Opportunities for Industrial Uptake in Trending IT Fields (SAGOIT-IT), фінансованому компанією ТОВ ГРУПБВТ. Строки виконання проєкту: 09.2020 – 01.2021. Науковий керівник: к.ф.-м.н., доц. Єрмолаєв В. А. Участь здобувача – виконавець.

6. Наукова новизна, теоретичне та практичне значення результатів дисертації

Наукова новизна результатів дослідження полягає у тому, що робота є першим у вітчизняній галузі комп'ютерних наук комплексним дослідженням, присвяченим розробці комплексного автоматизованого методу виявлення та вимірювання термінологічного насичення в колекціях професійних документів, таких як наукові публікації, для подальшого здобуття репрезентативних наборів термінів, що повно характеризують професійну предметну область. У результаті проведеного дослідження:

Уперше:

1. 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 Гільбертовому просторі колекцій документів, що є підмножинами повної колекції. Доведено теорему існування термінологічного насичення, що надає достатні умови існування. Розроблено алгоритми та програмне забезпечення для імплементації цього обчислювального методу.

Дістали подальшого розвитку:

2. Формальне визначення **міри термінологічної різниці** між двома наборами термінів зі значеннями оцінок їх значущості на базі (Tatarintseva et al. 2013) – міра формально визначена як різновид відстані Манхеттен (Goma and Fahmy 2013); формально доведено її метричні властивості.

3. 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 за (Frantzi and Ananiadou 1999) – тобто модифікація методу не зменшує якість здобуття термінів. Показано, що модифікований таким чином метод є коректним, ефективним (за показником часу виконання), таким, що очевидним чином розпаралелюється, та результативним з точки зору масштабованості для обробки текстів будь-якого великого обсягу.

Удосконалено:

4. Розроблений **обчислювальний конвеєр виявлення, вимірювання та аналізу термінологічного насичення**, що використовує розроблений комплексний метод виявлення та вимірювання термінологічного насичення та поліпшений метод автоматичного здобуття термінів з використанням злитих часткових C-value шляхом: залучення обчислювального методу для **відбору релевантних документів** до інкрементів колекції (Dobrovolskyi and Keberle 2018); використання розробленої техніки та алгоритмів **групування частково подібних термінів; впорядкування документів** для формування інкрементів колекції за зменшенням частоти цитування документів. За рахунок цих вдосконалень було підвищено якість та ефективність розробленого комплексного обчислювального методу.

7. Практичне значення результатів дисертації

Практичне значення дисертаційної роботи є суттєвим і підтверджується довідками про впровадження у промисловість та навчальний процес.

Промислове впровадження науково-технічного доробку дисертації здійснено у межах проекту «SAGOIT-IT: Strategic Analysis of R&D Gaps and Opportunities for Industrial Uptake in Trending IT Fields» компанії ТОВ ГРУПБВТ. У цьому проекті розроблений обчислювальний конвеєр було використано для перевірки прогнозу Гартнер щодо перспектив впровадження технології генеративних змагальних мереж поглибленого навчання в промисловість, шляхом аналізу термінологічного насичення.

Впровадження науково-технічного доробку дисертації у навчальний процес здійснено у рамках міжнародної магістерської програми з комп'ютерних наук та науки про дані Українського католицького університету. У цій програмі розроблений обчислювальний конвеєр було використано як програмний інструмент, що був рекомендований студентам на курсі «Академічне письмо» для відбору репрезентативного набору релевантних літературних джерел для написання огляду і аналізу сучасного стану досліджень за темами їх магістерських робіт.

8. Публікації, що висвітлюють основні результати дисертації, та особистий внесок здобувача

Наукові положення і результати, що представлені в дисертаційній роботі, отримані здобувачем особисто. Наукові публікації результатів дисертаційної роботи, написані у співавторстві. Нижче наведені ці публікації та вказано особистий внесок здобувача і розділи дисертації, що висвітлюються цими публікаціями:

1. **Kosa, V.**, Chugunenko, A., Yuschenko, E., Badenes-Olmedo, C., Ermolayev, V., Birukou, A.: Semantic saturation in retrospective text document collections. CEUR-WS, vol. 1851, 1–8 (2017) ISSN: 1613-0073. SCOPUS

Особистий внесок здобувача: план дисертаційного проекту; аналіз сучасного стану досліджень; візія підходу щодо виявлення та вимірювання термінологічного насичення в колекціях професійних документів; зібрані та передоброблені колекції документів.

Розділи дисертації: Вступ, 1.12, 1.13, 4.1, 4.2.

2. **Kosa, V.**, Chaves-Fraga, D., Naumenko, D., Yuschenko, E., Badenes-Olmedo, C., Ermolayev, V., Birukou, A.: Cross-evaluation of automated term extraction tools by measuring terminological saturation. Revised selected papers of ICTERI 2017. Cham, Germany: Springer-Verlag, CCIS vol. 826, 135–163 (2018) doi: 10.1007/978-3-319-76168-8_7, ISSN: 1865-0929. SCOPUS

Особистий внесок здобувача: збір та підготовка документів для синтетичних колекцій; експериментальна крос-оцінка програмних засобів для здобуття термінів; вибір базового методу та програмного засобу для автоматизованого здобуття термінів за результатами крос-оцінки

Розділи дисертації: 1.8, 1.9, 1.12, 1.13, 3.1 – 3.3, 3.5, 3.7, 3.8, 4.1 – 4.4.

3. Chugunenko, A., **Kosa, V.**, Popov, R., Chaves-Fraga, D., Ermolayev, V.: Refining terminological saturation using string similarity measures. CEUR-WS vol. 2105 3–18 (2018) ISSN: 1613-0073. SCOPUS

4. **Kosa, V.**, Chaves-Fraga, D., Keberle, N., Birukou, A.: Similar terms grouping yields faster terminological saturation. Revised selected papers of ICTERI 2018. Cham, Germany: Springer-Verlag, CCIS vol. 1007, 43–70. (2019) doi: 10.1007/978-3-030-13929-2_3, ISSN: 1865-0929. SCOPUS

Особистий внесок здобувача: алгоритми мір подібності, пороги подібності, алгоритм групування термінів **STG**, вдосконалений алгоритм **R-THD**;

експериментальна оцінка впливу групування подібних термінів на термінологічне насичення

Розділи дисертації: 1.10, 1.13, 3.6, 3.8, 4.1, 4.6.

5. **Kosa, V.**, Chaves-Fraga, D., Dobrovolskiy, H., Fedorenko, E., Ermolayev, V.: Optimizing automated term extraction for terminological saturation measurement. CEUR-WS, vol. 2387, 1–16 (2019) ISSN: 1613-0073. SCOPUS

Особистий внесок здобувача: оптимізований метод та алгоритми для обчислення злитих часткових C-value; доведення теореми про тотожність MPCV та C-value; експериментальна перевірка коректності методу

Розділи дисертації: 1.1 – 1.6, 1.11 – 1.13, 2.6, 3.3, 3.4, 3.8, 4.1, 4.7.

6. **Kosa, V.**, Chaves-Fraga, D., Dobrovolskiy, H., Ermolayev, V.: Optimized term extraction method based on computing merged partial C-values. Revised selected papers of ICTERI 2019. Cham, Germany: Springer-Verlag, CCIS vol. 1175, 24–49. (2020) doi: 10.1007/978-3-030-39459-2_2, ISSN: 1865-0929. SCOPUS

Особистий внесок здобувача: експериментальна перевірка незалежності від домену та результативності методу для колекції великого промислового обсягу

Розділи дисертації: 1.1 – 1.6, 1.11 – 1.13, 2.6, 3.3, 3.4, 3.8, 4.1, 4.7.

7. **Kosa, V.**, Ermolayev, V.: Toward a theoretical framework of terminological saturation for ontology learning from texts. CEUR-WS vol. 2566, 40–51 (2020) ISSN: 1613-0073. SCOPUS

Особистий внесок здобувача: формулювання ключових тверджень формального фреймворку про метричні властивості функції *thd* та достатні умови існування термінологічного насичення

Розділи дисертації: 2.1 – 2.5.

Таким чином, можна зазначити, що:

- 1) Здобувачем опубліковано сім наукових робіт, що висвітлюють основний зміст дисертації, усі з яких опубліковано у міжнародних періодичних виданнях з ISSN, що проіндексовані у базі даних SCOPUS.
- 2) Ці публікації достатньо повно розкривають основний зміст дисертації та відповідають умовам зарахування їх за темою дисертації відповідно пункту 11 Порядку проведення експерименту з присудження ступеня доктора філософії (Постанова КМУ № 167 від 6.03.2019 р. із змінами згідно з Постановою КМУ № 979 від 21.10.2020 р.).

9. 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.

Дисертацію оформлено у повній 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).

Дисертація написана грамотною українською мовою. Стиль викладення матеріалу відповідає прийнятому в науковій літературі з комп'ютерних наук, та характеризується точністю, логічністю, зрозумілістю, зв'язністю, цілісністю та завершеністю.

ВИСНОВОК

Ознайомившись із дисертацією Коси Вікторії Вікторівни «Метод експериментального дослідження термінологічного насичення в колекціях документів для здобуття знань» та науковими публікаціями, у яких висвітлені основні наукові результати дисертації, а також взявши до уваги підсумки фахового семінару, вважаємо, що:

1. Дисертація Коси Вікторії Вікторівни «Метод експериментального дослідження термінологічного насичення в колекціях документів для здобуття знань» є фундаментальним науковим дослідженням з актуальних питань, характеризується єдністю змісту, містить наукові результати, яким властива наукова новизна, теоретичне та практичне значення, а отже, свідчить про істотний особистий внесок здобувача у розвиток комп'ютерних наук.

2. Дисертація Коси Вікторії Вікторівни «Метод експериментального дослідження термінологічного насичення в колекціях документів для здобуття знань» може бути рекомендована до захисту на здобуття наукового ступеня доктора філософії за спеціальністю 122 – «Комп'ютерні науки» (галузь знань 12 – «Інформаційні технології») у разовій спеціалізованій вченій раді.

Рецензент:

професор кафедри
економічної кібернетики
Запорізького національного
університету,
д.ф.-м.н., професор

І. В. Козін

Рецензент:

професор кафедри
програмної інженерії
Запорізького національного
університету, д.т.н., доцент

С. В. Чопоров

Підпис
засвідчую

Чопорова С.В.
Козіна І.В.

НАЧАЛЬНИК
ВІДДІЛУ КАДРІВ



І. В. Козін