

АНОТАЦІЯ

Коса В. В. Метод експериментального дослідження термінологічного насичення в колекціях документів для здобуття знань. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 122 - Комп'ютерні науки. Запорізький національний університет Міністерства освіти і науки України, Запоріжжя, 2021.

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.

У вступі (розділ 1) 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. Приводяться відомості щодо структури та обсягу дисертаційної роботи.

У розділі 2 розглянуто та проаналізовано сучасний стан досліджень за тематикою роботи. Це допомогло розробити підхід до моделювання процесу термінологічного

насичення, виявлення та вимірювання результатів цього процесу. Розділ починається з викладу методології, яку було використано для відбору літературних джерел для нашого систематичного огляду і аналізу (розділ 2.1). Далі, у розділах 2.2 - 2.11 розглянуто та проаналізовано сучасний стан досліджень у релевантних наукових галузях, починаючи з розробки онтологій (Ontology Engineering) та вивчення онтологій (Ontology Learning) і закінчуючи якісними дослідженнями (Qualitative Research). У розділі 2.12 резюмовано виявлені прогалини у сучасному стані досліджень та мотивацію щодо зменшення цих прогалини. Базуючись на виявлених прогалинах, у розділі 2.13 запропоновано «дорожню карту» для вирішення виявлених відкритих питань. У розділі запропоновано бачення та окреслено підхід для вирішення цих питань дослідження. Крім того, сформульовано питання дослідження, на які необхідно відповісти у рамках окресленого підходу, щоб отримати ефективний та результативний обчислювальний метод. На базі питань дослідження поставлено завдання дослідження, вирішення яких має призвести до досягнення його мети. У розділі 2.14 підсумовано представлені результати огляду та аналізу сучасного стану досліджень.

Розділ 3 сфокусовано на виконанні першого завдання дослідження – розробці формального фреймворку для методу виявлення та вимірювання термінологічного насичення. Розділ починається з базових відомостей (розділ 3.1), що містять визначення, необхідні для окреслення фокусу обраного формального теоретичного підходу. Далі представлено гіпотези дослідження, які необхідно перевірити в контексті теоретичного підходу. Ці гіпотези сформульовано на основі питань дослідження, що згруповані у першому завданні роботи (розділ 2.13). Розділи 3.3 та 3.4 зосереджено на формальному введенні функції термінологічної різниці (*thd*) та доведенні її метричних властивостей у просторі усіх можливих колекцій документів для домену. У розділі 3.5 сформульовано та доведено теорему, що окреслює достатні умови існування термінологічного насичення. На додаток, у розділі 3.6, досліджено, чи можна оптимізувати метод вимірювання та виявлення термінологічного насичення з точки зору зменшення часу на обчислення.

У результаті, представлено розроблений вдосконалений метод, що використовує розділення (partitioning) колекції документів на частки, що не перетинаються. Доведено, що вдосконалений метод дає ті ж самі значення вимірювань *thd*, але вимагає значно менше часу для обчислень і не є обмеженим за обсягом колекції. Крім того, обґрунтовано, що вдосконалений метод можна розпаралелити. Отже, на додаток до поліпшеної ефективності, вдосконалений метод, на відміну від базового (розділи 3.1 – 3.5), є результативним при обробці великих колекцій документів, що мають реальні промислові обсяги.

У розділі 4 виконується друге завдання дослідження – розробка нових або вдосконалення раніше розроблених алгоритмів, що матеріалізують розроблений формальний фреймворк (розділ 3) у вигляді обчислювального методу. У розділі 4.1 представлено конвеєр обробки колекцій документів. Як подання високого рівня, представлено робочий процес (workflow) для вимірювання термінологічного насичення. Подальша деталізація цього процесу наведена у функціональній блок-схемі, яка розкриває модульну структуру нашого набору алгоритмів і залежності між модулями. Алгоритми, що є функціональним змістом модулів, представлено наступним чином. У розділі 4.2 представлені алгоритми, які розроблено для інструментальної підготовки даних. Вони включають генерацію каталогу колекції документів та завантаження загальнодоступних повнотекстових документів. У розділі 4.3 представлені алгоритми, розроблені для фази перед-обробки даних у робочому процесі. Вони призначені для перетворення PDF у плоский текст (plain text) та генерування наборів даних на основі визначених параметрів конфігурації. Оптимізований алгоритм здобуття термінів для статистичної частини конвеєру та алгоритм для обчислення об'єднаних часткових C-value детально описані у розділі 4.4. Базовий алгоритм для обчислення термінологічної різниці між двома наборами термінів представлено у розділі 4.5. У розділі 4.6 цей базовий алгоритм вдосконалюється шляхом врахування розробленої техніки групування термінів та алгоритмів вимірювання подібності символічних строк. Алгоритм видалення регулярного накопиченого шуму з наборів термінів наведено у розділі 4.7. У розділі 4.8 описується реалізація розроблених алгоритмів у програмному забезпеченні та

наводиться посилання на це програмне забезпечення, що є загальнодоступним для використання у академічних дослідженнях.

Завданням розділу 5 є третє завдання дослідження – експериментальна оцінка розробленого методу (розділ 3), матеріалізованого у наборі алгоритмів (розділ 4), для виявлення та вимірювання термінологічного насичення. У розділі 5.1 сформульовані завдання експериментального дослідження. План експериментів викладено у розділі 5.2. У розділах 5.3-5.7 повідомляються та обговорюються результати проведених експериментів. Як пояснюється у розділі 5, результати доводять правильність, незалежність від домену, ефективність та масштабованість розробленого обчислювального методу, набору алгоритмів та програмного забезпечення.

Розділ 6 виконує четверте завдання роботи – презентує досвід використання та візію того, як представлений науковий доробок доцільно впроваджувати в академічну та промислову практики. У розділі 6.1, представлено досвід використання розробленого програмного забезпечення в промисловому проєкті для крос-перевірки прогнозу Гартнер (Gartner) про тенденції впровадження технологій на прикладі технології генеративних недружніх мереж (Generative Adversarial Networks) у поглибленому машинному навчанні (Deep Learning). У розділі 6.2 повідомлено про академічне використання нашого методу та програмного забезпечення для проведення пошукових досліджень літератури студентами магістратури з метою написання оглядів релевантних джерел для своїх магістерських робіт. У розділі 6.3 узагальнено досвід практичного використання результатів роботи, представлений у розділах 6.1 та 6.2, шляхом висвітлення потенційних переваг розробленого методу та програмного забезпечення для промислових користувачів. У розділі 6.4, ми представляємо потенційні бізнес-сценарії щодо застосування результатів роботи у галузі наукового видавництва. Досвід та перспективи практичного використання результатів роботи підсумовано у розділі 6.5.

У розділі 7 представлені загальні висновки по роботі та плани щодо подальшого розвитку науково-технічного доробку.

Ключові слова: предметна область, колекція документів, термінологічне насичення, виявлення термінологічного насичення, вимірювання термінологічного насичення, послідовне наближення, обчислювальний метод, набір алгоритмів, ефективність, результативність.

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА

Наукові праці, в яких опубліковано основні наукові результати дисертації:

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