УДК 004.912:004.8 **doi: 10.26906/SUNZ.2025.3.96-100**

О. І. Іванюк

Український державний університет залізничного транспорту, Харків, Україна

МЕТОДОЛОГІЯ АВТОМАТИЗОВАНОЇ ОЦІНКИ ТЕМАТИЧНОЇ СПОРІДНЕНОСТІ НАУКОВИХ ПУБЛІКАЦІЙ НА ОСНОВІ СЕМАНТИЧНОГО АНАЛІЗУ

Анотація. Актуальність. Автоматизована оцінка тематичної спорідненості між дисертаційними дослідженнями та профілями потенційних експертів потрібна для прозорого й відтворюваного добору офіційних опонентів і складу разових рад; відкриті дані NAQA.Svr роблять таку оцінку технічно можливою. **Об'єкт дослідження**: методи та засоби побудови семантичних профілів дисертацій і науковців та їх ранжування у спільному векторному просторі. **Мета статті**: розробити та емпірично перевірити методологію автоматизованої оцінки тематичної спорідненості наукових публікацій на основі семантичного аналізу. **Результати дослідження**. Сформовано корпус із 259 дисертацій, 662 профілів науковців і 3345 публікацій; тексти нормалізовано, назви публікацій стандартизовано засобами великої мовної моделі. Семантичні подання отримано за допомогою моделі OpenAI. Порівнювалися два варіанти подання дисертацій (повний опис; лише ключові слова) та науковців (публікації; ключові слова). Якість оцінювалася за фактичними призначеннями опонентів, використовуючи метрики NDCG, Hit@k і NDCG-lift. Найкращі результати стабільно показала комбінація «дисертація за ключовими словами» та «науковець за публікаціями», яка демонструє найвищу точність у верхніх позиціях і найбільший виграш відносно випадкового відбору. **Висновки.** Поєднання профілю дисертації за ключовими словами з профілем науковця, агрегованим за публікаціями, забезпечує найкращий баланс точності та стійкості і є практично доцільним для попереднього добору опонентів. Запропонована методологія відтворювана, масштабована й узгоджена з відкритими даними; перспективні напрями – урахування часової ваги публікацій, двомовності термінів та процедурних обмежень під час остаточного призначення.

Ключові слова: тематична спорідненість; ембедінги; косинусна подібність; призначення рецензентів; дисертації; профілі науковців; OpenAI text-embedding-3-large; UMAP; DBSCAN; NDCG; Hit@k; рекомендаційні системи; текстова аналітика, машинне навчання.

Вступ

Постановка проблеми. Оцінка тематичної спорідненості між науковими текстами та профілями дослідників є базовим інструментом для добору рецензентів, формування експертних рад і призначення наукових керівників.

У вітчизняній практиці цю потребу прямо закріплено в регуляторних вимогах до разових спеціалізованих вчених рад: члени рад мають бути компетентними саме за тематикою дисертації, що операційно визначається наявністю не менше трьох публікацій за темою дослідження, опублікованих протягом останніх п'яти років до дня утворення разової ради і після присудження власного ступеня (кандидата наук/доктора філософії) [1]. Процедурна прозорість і перевірюваність такої відповідності підтримуються інформаційною системою NAQA.Svr, що акумулює дані про склад разових рад і матеріали атестаційних процедур [2].

У сфері акредитації освітньо-наукових програм вимога тематичної відповідності також має нормативне відображення: методичні орієнтири для програм третього рівня роблять акцент на «навчанні через дослідження» та узгодженості дослідницьких тем аспірантів із напрямами наукової діяльності їхніх керівників (підтверджується релевантними публікаціями), що відбивається у положеннях МОН про акредитацію та рекомендаціях НАЗЯВО щодо застосування критеріїв якості [3–5]. Таким чином, нормативна база послідовно вимагає змістовної (тематичної) відповідності і для складу рад, і для кадрового забезпечення PhD-програм.

На практиці ручна перевірка тематичної відповідності у разових радах і при доборі опонентів є трудомісткою, суб'єктивною та важко масштабованою: потрібно зіставляти зміст дисертацій із публікаційними профілями великої кількості потенційних експертів, враховуючи нормативні обмеження (компетентність саме за тематикою, публікації за останні п'ять років тощо) [1]. Запроваджена прозорість і відкритість процедур через інформаційну систему NAQA.Svr створює технічні передумови для автоматизованих інструментів попереднього тематичного відбору, які підвищують відтворюваність і швидкодію процесу, зменшують ризики упередженості та пропуску релевантних фахівців, а також забезпечують аудитовність прийнятих рішень у відповідності до вимог Порядку [1, 2].

Аналіз останніх досліджень і публікацій. У міжнародній літературі завдання автоматизованого призначення рецензентів (Reviewer Assignment Problem, RAP) зазвичай подають як триетапну схему:

I – побудова профілів заявки та рецензентів;

II – обчислення міри відповідності;

III – оптимізація призначень з урахуванням обмежень (баланс навантаження, конфлікти інтересів тощо) [6, 7].

На практиці застосовують як класичні системи (наприклад, Toronto Paper Matching System, SubSift), так і новіші підходи, що вдосконалюють як блок обчислення відповідності, так і оптимізаційні процедури призначення [8, 9].

Методи оцінювання тематичної спорідненості еволюціонували від лексичних (показник TF-IDF,

ключові слова) та латентних моделей (LSA, LDA) до семантичних векторних подань і контекстних мовних моделей (word2vec, BERT, SciBERT тощо). Огляди фіксують, що поєднання лексичних і семантичних ознак системно підвищує якість відповідності та стійкість до варіативності термінів [6–8, 10, 11].

Зокрема, [8] показано переваги ітеративної моделі WSIM, яка поєднує мовні та тематичні ознаки; у [11] продемонстровано ефективність тематичної експертизи для співставлення «стаття-рецензент»; застосування векторних подань (word2vec) також пропонувалося у фреймворках контентного зіставлення [10].

В українському контексті професором Сергієм Штовбою та доктором філософії Миколою Петриченком запропоновано підхід експрес-підбору для разових рад, що фокусується на швидкому звуженні великого пулу кандидатів за тематикою з наступним «тонким» добором. Автори підкреслюють важливість командного покриття тематики дисертації (щоб колектив членів ради сумарно охоплював усі аспекти теми) та демонструють покращення тематичної відповідності складу рад у середньому на 13–34% порівняно з ручним добором [12].

Метою роботи є розробка методології автоматизованої оцінки тематичної спорідненості наукових публікацій на основі семантичного аналізу, що реалізує етап II RAP (обчислення міри відповідності між дисертаційними дослідженнями та публікаційними профілями потенційних членів рад) із залученням універсальних моделей ембедингів OpenAI та відкритих даних інформаційної системи NAQA.Svr.

Методи. Використано відкриті дані NAQA.Svr, нормалізацію текстів, семантичні векторні подання (OpenAI text-embedding-3-large), проєкцію UMAP, кластеризацію DBSCAN, косинусну подібність для обчислення спорідненості та метрики NDCG@k, Hit@k і lift для оцінки якості.

Для мовного редагування рукопису застосовано асистента на основі штучного інтелекту ChatGPT (OpenAI).

Основний матеріал

Позначимо множину потенційних експертів (науковців) через $S = \{s_1, ..., s_{|S|}\}$. Кожному $s \in S$ відповідає скінченна множина пов'язаних текстових одиниць $U_s = \{u_1, ..., u_{|U_s|}\}$ (наприклад, заголовки публікацій та/або конкатеновані ключові слова з їхніх публікацій).

Множину дисертацій (або інших заявок) позначимо через $T = \{t_1, ..., t_{|T|}\}$, де кожному $t \in T$ відповідає множина текстових одиниць $V_t = \{v_1, ..., v_{|V_s|}\}$ (наприклад, теми, ключові слова, анотації).

Нехай $\phi : \text{Text} \to \mathbb{R}^d$ – фіксоване відображення тексту в евклідів простір $\mathbb{R}^d$ (семантичне векторне подання, ембединг).

Профіль науковця векторизується як:

$$e_s = \text{Agg}_S(\{\phi(u) : u \in U_s\}) \in \mathbb{R}^d, \quad (1)$$

де Agg_S – оператор агрегації множини векторів (наприклад, середнє, зважене середнє, робастні усереднення тощо).

Аналогічно, профіль дисертації визначається як:

$$e_t = \text{Agg}_T(\{\phi(v) : v \in V_t\}) \in \mathbb{R}^d, \quad (2)$$

де Agg_T – оператор агрегації для дисертаційних одиниць.

Такий e_s інтерпретується як «тематичний центроїд» науковця.

Міра тематичної спорідненості між дисертацією t та науковцем s задається косинусною подібністю:

$$\text{sim}(t, s) = \frac{e_t^T e_s}{\| e_t \|_2 \| e_s \|_2} \in [-1, 1], \quad (3)$$

а за потреби відстань визначається як $d(t, s) = 1 - \text{sim}(t, s)$.

Для даної дисертації t розглядається множина допустимих кандидатів $S_t \subseteq S$ (після застосування регуляторних та процедурних обмежень щодо конфлікту інтересів чи ролі у раді). Визначається функція оцінювання $f_t(s) = \text{sim}(t, s)$ та індукована нею функція ранжування π_t, що впорядковує S_t за спаданням $f_t(s)$. Рекомендована множина k-кращих визначається як:

$$R_k(t) = \{s \in S_t : \text{rank}_{\pi_t}(s) \le k\}, \quad (4)$$

де $\text{rank}_{\pi_t}(s)$ – порядковий номер s у списку, 1 – найвища релевантність.

Для практичної реалізації дані збиралися з відкритих записів інформаційної системи NAQA.Svr у межах спеціальностей 121–126 за період червень 2022 – грудень 2024. Отримання відомостей здійснювалося шляхом автоматизованого доступу до вебсторінок з подальшим HTML-парсингом табличних блоків та вмісту модальних вікон зі списками публікацій членів ради.

Для кожного захисту фіксувалися:

– тема дисертації,

– анотація,

– ключові слова;

– персональні переліки публікацій кожного члена ради (ПІБ, роль у раді, афіляція, рік і ключові слова публікацій).

Первинні тексти проходили нормалізацію:

– приведення до нижнього регістру,

– уніфікація лапок і тире,

– видалення подвоєних пробілів;

– у формулюваннях тем прибиралися фінальні крапки;

– ключові слова очищувалися від технічних артефактів і зводилися до єдиного формату (нижній регістр, усунення порожніх елементів).

Назви публікацій уніфікувалися за допомогою автоматичного вилучення стандартної назви з бібліографічних або вільно сформованих описів із вико-

ристанням великої мовної моделі (GPT-4o-mini) через API OpenAI (інструктивний запит із вимогою повернути лише назву). Це дозволило усунути дублікати і зменшити шум у подальшому визначенні тематичної спорідненості.

У результаті сформовано корпус із 259 дисертацій (тема, ключові слова та анотація; спеціальні/закриті теми виключено), 662 унікальних науковців (голови, рецензенти, опоненти) і 3345 унікальних публікацій.

Семантичні подання текстів формувалися в єдиному просторі розмірності 1024 за допомогою універсальної моделі ембедингів OpenAI (text-embedding-3-large).

Для дисертацій було реалізовано два варіанти профілювання:

1) Д(ТКА) – вектор, отриманий із конкатенованого опису «тема + ключові слова + анотація»;

2) Д(КС) – вектор, побудований лише на основі ключових слів.

Аналогічно, для науковців сформовано два варіанти профілю:

1) Н(ПУБ) – агрегований вектор, що дорівнює середньому L2-нормованих ембедингів усіх публікацій науковця, де ембединг кожної публікації будувався з конкатенованого опису «назва + ключові слова»;

2) Н(КС) – ембединг конкатенованого списку всіх ключових слів публікацій науковця.

Таким чином, утворилося чотири комбінації для порівняльного аналізу. Єдиний простір і однакова розмірність забезпечують коректність використання косинусної подібності та роблять результати сумірними між різними комбінаціями профілів.

Візуальна перевірка простору ембедингів виконувалася за допомогою проєкції UMAP у двовимірний простір з подальшим виділенням щільнісних груп методом DBSCAN.

На площині відображено окремо дисертації та профілі науковців; кластерні мітки присвоювалися автоматично, «шумові» точки відсікалися як ті, що не входять до жодного щільного скупчення.

Ілюстрації показують виразні тематичні групи й їхні перетини: наприклад:

– кластер «кібербезпека/інформаційна безпека»,

– великі скупчення «машинне/глибоке навчання»,

– спеціалізовані кластери «метод скінченних елементів/обчислювальна механіка», «управління проєктами», «БПЛА» тощо.

Для профілів науковців Н(КС) (рис. 1) спостерігається тонший поділ на теми (24 кластери, «шум» ≈ 11.8%), ніж для профілів Н(ПУБ) (17 кластерів, «шум» ≈9.4%).

Для профілів дисертацій Д(ТКА) виділено 15 кластерів («шум» ≈11.2%), тоді як профілювання Д(КС) (рис. 2) – 13 кластерів із найменшою часткою «шуму» (≈4.2%).

Така картина узгоджується з інтуїцією: ключові слова забезпечують чіткіші тематичні межі, тоді як повний опис краще відтворює змістові перетини між темами.

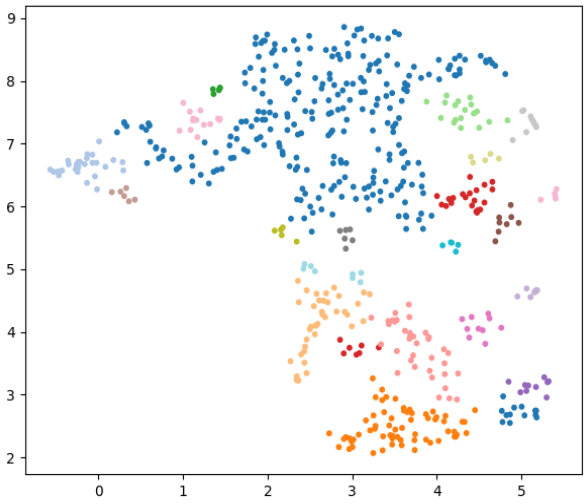

Рис. 1. UMAP+DBSCAN для профілів науковців Н(КС)

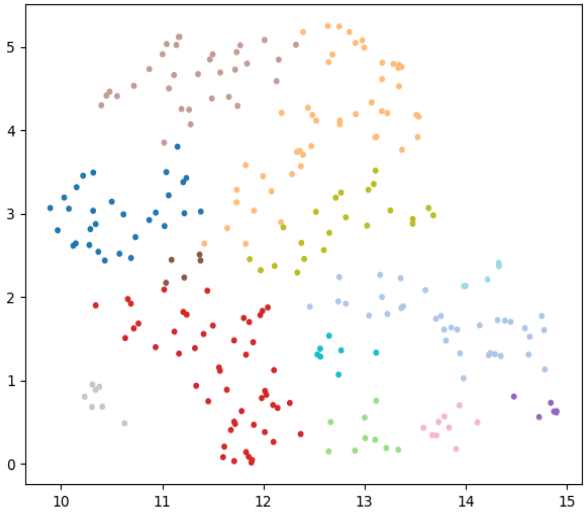

Рис. 2. UMAP+DBSCAN для профілів дисертацій Д(КС)

Оцінювання якості визначення тематичної спорідненості виконано на сформованому корпусі. Еталоном релевантності задавалися фактично призначені офіційні опоненти. Для кожної дисертації множина кандидатів формувалася з усіх профілів науковців за винятком інших учасників відповідної разової ради, після чого чотири комбінації профілів {Д(ТКА), Д(КС)}×{Н(ПУБ), Н(КС)} порівнювалися за косинусною подібністю (3).

Ранжування оцінювалося за метрикою NDCG@k (чутлива до позицій співпадінь у верхній частині списку) та Hit@k (частка випадків, коли хоча б один опонент потрапив у k-кращих). Додатково розглядався коефіцієнт підсилення NDCG (NDCG-lift), тобто відношення значення метрики моделі до її випадкового еталона; невизначеність оцінювалася бутстреп-методом з 95% довірчими інтервалами. Графік із коефіцієнтом підсилення (рис. 3) демонструє стійку перевагу комбінації Д(КС)↔Н(ПУБ): для малих k вона суттєво випереджає альтернативи, а зі збільшенням k зберігає найвищі значення при помірній ширині довірчих інтервалів.

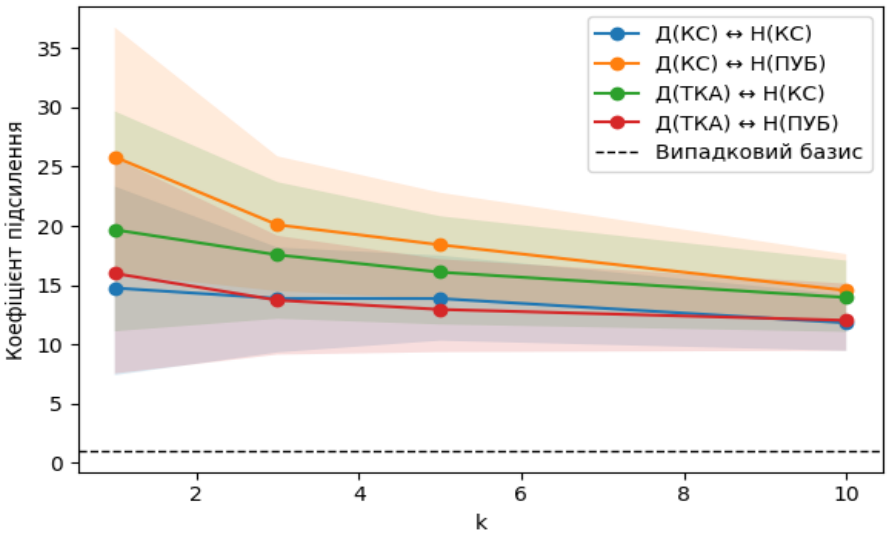

Рис. 3. Криві коефіцієнта підсилення NDCG для чотирьох комбінацій профілів дисертацій та науковців із 95% довірчими інтервалами

Друге місце стабільно посідає Д(ТКА)↔Н(КС); далі йде Д(ТКА)↔Н(ПУБ), тоді як Д(КС)↔Н(КС) показує найнижчі. Спадання коефіцієнта підсилення NDCG (рис. 1) із ростом k є очікуваним: базовий випадковий рівень поліпшується швидше, тож відносна перевага моделі зменшується.

Таблиця 1 – **Середні значення метрик NDCG та Hit@k, а також коефіцієнти підсилення NDCG відносно випадкового базису.**

Комбінація	k	NDCG	NDCG lift	Hit@k
Д(ТКА) ↔ Н(ПУБ)	1	0.05	15.98	0.05
	3	0.055	13.72	0.12
	5	0.072	12.94	0.174
	10	0.103	12.02	0.301
Д(ТКА) ↔ Н(КС)	1	0.062	19.67	0.062
	3	0.07	17.55	0.147
	5	0.089	16.09	0.208
	10	0.119	13.95	0.347
Д(КС) ↔ Н(ПУБ)	1	0.081	25.82	0.081
	3	0.08	20.09	0.17
	5	0.102	18.40	0.232
	10	0.124	14.54	0.317
Д(КС) ↔ Н(КС)	1	0.046	14.75	0.046
	3	0.056	13.87	0.127
	5	0.077	13.86	0.201
	10	0.101	11.81	0.309

Табл. 1 зі значеннями NDCG, Hit@k та коефіцієнтом підсилення NDCG підтверджує ці спостереження чисельно та дає інтуїцію щодо якості для різних довжин списку рекомендацій.

Висновки

Розроблена методологія автоматизованого зіставлення дисертацій з профілями потенційних експертів у спільному семантичному просторі показала стабільну перевагу комбінації «дисертація за ключовими словами» у поєднанні з «профілем науковця, усередненим за публікаціями».

Такий результат узгоджується з інтуїцією: ключові слова концентрують тематичне ядро дослідження, а усереднення публікаційного профілю науковця надає стійку оцінку його наукового поля. Візуалізації (UMAP+DBSCAN) додатково підтверджують наявність виразних тематичних скупчень і помірних перетинів між темами.

Практична цінність підходу полягає у прозорому та відтворюваному попередньому доборі офіційних опонентів і формуванні складу разових рад, а також у задачах узгодження тем аспірантів із керівниками та пошуку зовнішніх рецензентів для конкурсних проєктів.

Метод легко інтегрується в наявні робочі процеси, забезпечуючи пояснювані рейтинги кандидатів і аудитованість рішень.

Подальший розвиток роботи вбачається:

– у запровадженні часових ваг для публікацій,

– використанні доменно-адаптованих моделей отримання ембедингів,

– застосуванні методів навчання для ранжування на історичних даних,

– інкорпорації графових сигналів (співавторство, інституційні зв’язки) та розробленні механізмів пояснюваності рекомендацій.

Такі розширення мають підвищити точність і стійкість без втрати прозорості та масштабованості запропонованого рішення.

СПИСОК ЛІТЕРАТУРИ

1. Кабінет Міністрів України. (2022). Про затвердження Порядку присудження ступеня доктора філософії та скасування рішення разової спеціалізованої вченої ради закладу вищої освіти, наукової установи про присудження ступеня доктора філософії. Available at: https://zakon.rada.gov.ua/laws/show/44-2022-%D0%BF#Text
2. Національне агентство із забезпечення якості вищої освіти. (2024). Порядок функціонування інформаційної системи «NAQA.Svr» (затверджено 30.08.2022; зі змінами від 19.11.2024). Available at: https://naqa.gov.ua/wp-content/uploads/2025/04...8F-NAQA.Svr_22.04.2025.pdf
3. Національне агентство із забезпечення якості вищої освіти. (2020). Рекомендації для експертів Національного агентства стосовно акредитації освітніх програм третього рівня вищої освіти (додаток до «Методичних рекомендацій для експертів Національного агентства щодо застосування Критеріїв оцінювання якості освітньої програми»). Available at: https://naqa.gov.ua/wp-content/uploads/2020/02....pdf
4. Міністерство освіти і науки України. (2024). Про затвердження Положення про акредитацію освітніх програм, за якими здійснюється підготовка здобувачів вищої освіти (Наказ № 686; зареєстровано в Міністерстві юстиції України 04.07.2024 № 1013/42358). Available at: https://zakon.rada.gov.ua/laws/show/z1013-24#Text
5. Національне агентство із забезпечення якості вищої освіти. (2019). Методичні рекомендації для експертів Національного агентства щодо застосування Критеріїв оцінювання якості освітньої програми. Available at: https://bit.ly/4gG5Yki
6. Zhao, X., & Zhang, Y. (2022). Reviewer assignment algorithms for peer review automation: A survey. Information Processing & Management, 59(5), 103028. https://doi.org/10.1016/j.ipm.2022.103028
7. Aksoy, M., Yanik, S., & Amasyali, M. F. (2023). Reviewer assignment problem: A systematic review of the literature. *Journal of Artificial Intelligence Research*, 76, 761–827. https://doi.org/10.1613/JAIR.1.14318
8. Tan, S., Duan, Z., Zhao, S., Chen, J., & Zhang, Y. (2021). Improved reviewer assignment based on both word and semantic features. *Information Retrieval Journal*, 24(3), 175–204. https://doi.org/10.1007/s10791-021-09390-8
9. Price, S., & Flach, P. (2017). Computational support for academic peer review: A perspective from artificial intelligence. *Communications of the ACM*, 60(3), 70–79. https://doi.org/10.1145/2979672
10. Ogunleye, O., Ifebanjo, T., Abiodun, T. N., & Adebiyi, A. A. (2017). Proposed framework for a paper-reviewer assignment system using Word2Vec. *Proc. of the 4th Covenant University Conf. on E-Governance in Nigeria* (CUCEN 2017). Covenant University Repository. Available at: https://www.researchgate.net/publication/322356955_Covenant_University_E-Governance_Conference_CUCEN_2017
11. Mimno, D., & McCallum, A. (2007). Expertise modeling for matching papers with reviewers. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 500–509. https://doi.org/10.1145/1281192.1281247
12. Штовба, С. Д., Петричко, М. В. (2024). Експрес-підбір опонентів для разових рад із захисту PhD-дисертацій. *Ukrainian Journal of Information Systems and Data Science*, 2(1), 41–54. https://doi.org/10.31558/2786-9482.2024.1.4

Received (Надійшла) 29.05.2025
Accepted for publication (Прийнята до друку) 27.08.2025

ВІДОМОСТІ ПРО АВТОРІВ / ABOUT THE AUTHORS

Іванюк Олександр Ігорович – доктор філософії, доцент кафедри інформаційних технологій, Український державний університет залізничного транспорту, Харків, Україна;
Oleksandr Ivaniuk – PhD, Associate Professor of Department of Information Technology, Ukrainian State University of Railway Transport, Kharkiv, Ukraine;
e-mail: aleks.ivaniuk@gmail.com; ORCID Author ID: https://orcid.org/0000-0002-4007-2215;
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57203140184.

Methodology for automated assessment of thematic relatedness of scientific publications using semantic analysis

Oleksandr Ivaniuk

Abstract. Relevance. Automated assessment of thematic relatedness between doctoral theses and profiles of potential experts is needed for transparent and reproducible selection of official reviewers and composition of one-time specialized councils; the open NAQA.Svr data make such assessment technically feasible. **Object of research**: methods and tools for constructing semantic profiles of theses and researchers and for ranking them in a shared vector space. **Purpose of the article**. To develop and empirically validate a methodology for automated assessment of thematic relatedness of scientific publications based on semantic analysis. **Research results**. We assembled a corpus of 259 theses, 662 researcher profiles, and 3345 publications; texts were normalized, and publication titles were standardized using a large language model. Semantic representations were obtained with an OpenAI model. We compared two variants of thesis profiling (full description; keywords only) and two variants of researcher profiling (by publications; by keywords). Quality was evaluated against actual reviewer assignments using the metrics NDCG, Hit@k, and NDCG-lift. The best results were consistently achieved by the combination “thesis by keywords” and “researcher by publications,” which yields the highest top-rank accuracy and the largest gain over random selection. **Conclusions**. Combining a keyword-based thesis profile with a publication-aggregated researcher profile provides the best balance of accuracy and robustness and is practically suitable for preliminary reviewer selection. The proposed methodology is reproducible, scalable, and aligned with open data; promising directions include incorporating temporal weights of publications, handling bilingual terminology, and accounting for procedural constraints in final assignments.

Keywords: thematic relatedness; embeddings; cosine similarity; reviewer assignment; dissertations; researcher profiles; OpenAI text-embedding-3-large; UMAP; DBSCAN; NDCG; Hit@k; recommender systems; text analytics; machine learning.