УДК 001.02.3

Експрес-підбір опонентів для разових рад із захисту PhD-дисертацій

Сергій Штовба
професор, д-р техн. наук
ORCID: 0000-0003-1302-4899
s.shtovba@donnu.edu.ua

Донецький національний університет імені Василя Стуса

Микола Петричко
ORCID: 0000-0001-6836-7843
mpetrychko@vntu.edu.ua

Вінницький національний технічний університет

Ключові слова:
задача про призначення рецензентів;
експрес-підбір;
обробка природної мови;
категоризація;
дискретна оптимізація;
аналіз даних;
Dimensions.

Сьогодні ради із захисту PhD-дисертацій формують у ручному режимі. Це обумовлює як корупційні ризики, так і значні витрати часу на пошук та аналіз кандидатів з великими шансами пропустити кваліфікованих опонентів. Тому виникає зацікавленість у автоматизації формування разових рад для усунення зазначених ризиків впливу людського фактора. Стаття фокусується на експрес-підборі рад, коли потрібно сильно звузити великий список кандидатів. Подальший короткий список можна аналізувати або вручну, або передавати на процедуру тонкого підбору, яка є ресурсно-витратною і вимагає значно більшого об'єму початкової інформації. Пропонується метод призначення команди рецензентів за їх відповідністю тематиці дисертації, який, на відміну від ізольованого підбору кандидатів, враховує здатність саме колективу рецензентів спільно оцінити роботу за всіма аспектами її тематики. Метод є збалансованим за критеріями якості підбору і витратами ресурсів на пошук членів ради. Метод включає три етапи. На першому етапі здійснюється категоризація дисертації та потенційних членів ради шляхом представлення їх тематик векторами у просторі наукових спеціальностей з ANZSRC-2020. На другому етапі розраховується рівень відповідності кандидатів тематиці дисертації з урахуванням спорідненості наукових спеціальностей ANZSRC-2020. На третьому етапі підбирається склад ради, яка відповідає тематиці дисертації з максимально можливим ступенем. Для реалізації третього етапу запропоновано кілька алгоритмів оптимізації. Тестування алгоритмів на сформованому датасеті із 67 PhD-дисертацій показало, що найкращий баланс за критеріями якості підбору й витрат ресурсів на пошук колективу забезпечують жадібний алгоритм без елітизму та повний перебір на прорідженій множині кандидатів. Внаслідок оптимізації вдалося покращити склад разових рад у середньому на 13–34% залежно від типу використаного алгоритму.

DOI: 10.31558/2786-9482.2024.1.4

Вступ

В Україні PhD-дисертації захищають у разових радах. Разова рада складається із 5 науковців, які мають бути фахівцями з тематики дисертації. Голова ради та 1 або 2 рецензенти представляють заклад, у якому утворюється разова рада, а 2 чи 3 опоненти

запрошують з інших установ. Членів разової ради добирають вручну і затверджують її склад рішенням вченої ради закладу.

Ручне формування разової ради має кілька недоліків. По-перше, це корупційні ризики, коли раду формують виключно з дружніх осіб, які апріорі надають лише схвальні відгуки незалежно від результатів дисертації. В успішному захисті дисертації зацікавлені як особи, які підбирають членів разової ради, так і заклад, який її затверджує. За даними з інформаційної системи НАЗЯВО на поточний момент відбулося 5 324 успішні захисти дисертацій; випадків відмов присудження наукового ступеня немає. Інколи успішно захищають не лише відверто слабкі дисертації, але навіть явно псевдонаукові. Бували навіть випадки, коли здобувачі ледь-ледь розмовляли українською, але це не вплинуло на ухвалення позитивних рішень. По-друге, витрачається багато часу на ручний пошук та аналіз кандидатів у члени ради. Щомісяця в Україні формується приблизно 300 разових рад. Якщо припустити, що на пошук членів однієї ради в середньому витрачається 10 людино-годин, тоді за місяць набігає 3 000 людино-годин, що еквівалентно 18 ставкам. По-третє, сформований склад ради може не повністю відповідати тематиці дисертації через те, що когось із хороших кандидатів упустили під час ручного пошуку. Тому виникає зацікавленість у автоматизації формування разових рад для усунення зазначених ризиків впливу людського фактора.

Формування разових рад із захисту PhD-дисертації є однією із задач призначення рецензентів наукових творів, яку в англомовній літературі називають *Reviewer Assignment Problem* або *Paper-Reviewer Assignment*. Для специфічних задач призначення рецензентів використовують також терміни *Committee Review Assignment* та *Conference Paper Assignment Problem* [1]. Тут під призначенням рецензентів розуміється підбір усіх осіб, які оцінюють наукові твори. За термінологією цих задач твір, який оцінюється, називається заявкою. В контексті експертизи PhD-дисертації рецензентами є усі члени разової ради – і голова ради, і опоненти, і номінальні рецензенти.

Задача призначення рецензентів складається з трьох етапів [2, 3]: 1) пошук рецензентів і вибір методу представлення даних про рецензентів та заявки; 2) підрахунок схожості між заявкою та рецензентами; 3) розподіл заявок за рецензентами для максимізації агрегованої схожості за всіма призначеннями за деяких обмежень. Типовими обмеженнями є збалансованість навантажень рецензентів, врахування їх вподобань та запобігання конфлікту інтересів. У цій роботі вважається, що список потенційних рецензентів наявний.

Автоматичний підбір рецензентів передбачає доступність деякої початкової інформації про рецензентів та заявки. Структуровану сукупність такої інформації називають профілем рецензента та профілем заявки. Доволі часто для побудови профіля рецензента використовується така інформація про його статті: назва, анотація, ключові слова, повний текст, список посилань та список цитувань [4]. Профіль також може містити і особисті дані рецензента. Для створення профіля заявки найчастіше використовують анотацію, повний текст, ключові слова, назву роботи та галузь дослідження [4].

Побудова профілів заявки та рецензентів здійснюється за допомогою різноманітних методів обробки природної мови на основі мішка слів [4, 5, 6], прихованого семантичного аналізу [7, 8], тематичного моделювання [9, 10, 11], статичних мовних моделей з глибоким навчанням [11, 12, 13, 14, 15, 16] та контекстуальних моделей з глибоким навчанням [17, 18,

19, 20, 21]. Підходи до вирішення задачі автоматичного призначення рецензентів у більшості випадків потребують доволі великого обсягу початкової інформації про публікації рецензентів, їх взаємодію з іншими науковцями та аналогічну інформацію про авторів заявок. Оброблення такої інформації є витратним і не буде доцільним, якщо під кожну разову раду детально аналізувати тисячі кандидатів.

Ми фокусуємося на задачі експрес-підбору рецензентів, коли потрібно сильно скоротити довгий початковий список кандидатів. Подальший короткий список можна аналізувати вручну, або активувати процедуру тонкого підбору, яка є ресурсно-витратною і вимагає значно більшого об'єму початкової інформації, ніж це потрібно для експрес-підбору. Під час експрес-підбору враховується лише семантична схожість між заявками та рецензентами – відбувається підрахунок індексів схожості і далі призначається необхідна кількість рецензентів так, щоб забезпечити максимальну відповідність колективу рецензентів заявці за деяким критерієм. Розробка ефективного алгоритму експрес-формування разової ради для PhD-дисертацій, який є збалансованим за критеріями якості підбору та витрат ресурсів на формування колективу, і є ***метою*** цього дослідження.

Представлення даних про заявку та рецензентів

На першому етапі призначення рецензентів необхідно обрати початкові дані для прийняття рішення, а також метод їх представлення у векторній формі. У випадку заявки використовується список її ключових слів, а у випадку рецензента – список ключових слів, який отримується з доступних даних. У загальному випадку цей список ключових слів може бути як зі свіжих публікацій кандидата, так із його CV чи із профілю з деякого реєстру науковців. За другого випадку ключові слова або дослідницькі інтереси кандидат формує на власний розсуд, тобто вони представлені у довільній формі без прив'язки до будь-якого рубрикатора чи класифікатора. За поточної практики захисту вітчизняних PhD-дисертацій тематика члена спецради описується ключовими словами кількох його свіжих статей.

Початкові дані зазвичай обробляють з використанням статистичних моделей, тематичних моделей та моделей ембедингу. Деякі з них аналізують частоту появи слів у тексті, інші формують вектори представлення на основі співпояви слів. Зазвичай результуючі векторні представлення складно інтерпретувати. До того ж для отримання таких представлень необхідна велика кількість даних. Ми пропонуємо використовувати підхід з [22], за яким сукупність ключових слів категоризується і відображається як вектор у просторі наукових спеціальностей з Австралійсько-Новозеландської системи класифікації наук ANZSRC-2020. ANZSRC-2020 включає в себе 171 спеціальність із 22 галузі. Отже, кінцеве представлення профілів заявки та рецензента виглядає як розподіл над 171 науковими спеціальностями з ANZSRC-2020.

Щоб здійснити категоризацію, необхідно мати корпус розмічених статей, які приписані до однієї чи кількох наукових спеціальностей, та модель машинного навчання, яка за ключовими словами віднесе аналізований профіль до тих чи інших спеціальностей. Це потребує великої кількості людських ресурсів для розмітки статей та постійного оновлення даних. Ми пропонуємо використовувати інформаційні ресурси системи Dimensions, в якій понад 100 мільйонів публікацій вже категоризовано за ANZSRC-2020. За пошуковим

запитом у формі ключового слова Dimensions формує видачу, в якій зазначається, скільки публікацій із цим ключовим словом віднесено до кожної зі спеціальностей. Схематично ця процедура зображена на рис. 1. З нього також видно, що в базі розмічених документів стаття може категоризуватися до кількох спеціальностей, наприклад, *Стаття 1* віднесена до *Науки 1* та *Науки 2*. На основі такої видачі можна побудувати розподіл появи ключового слова в контексті різних наук. Наприклад, для ключового слова з рис. 1 розподіл появи матиме такий вигляд: *Наука 1* – 1 поява, *Наука 2* – 1 поява, *Наука 3* – 2 появи. На основі такого розподілу далі виконується категоризація ключового слова "*neural network*" у межах системи класифікації наук. Для категоризації множини ключових слів застосуємо алгоритм із [22], який опирається на ресурси та сервіси інформаційної системи Dimensions. Цей алгоритм враховує як появу ізольованих ключових слів із профілю рецензента чи заявки, так і спільну появу пар ключових слів. Алгоритм дає змогу відфільтрувати інформаційні шуми, що спричиненні як стоп-словами, так і рідкими ключовими словами, достовірність висновків за якими низька.

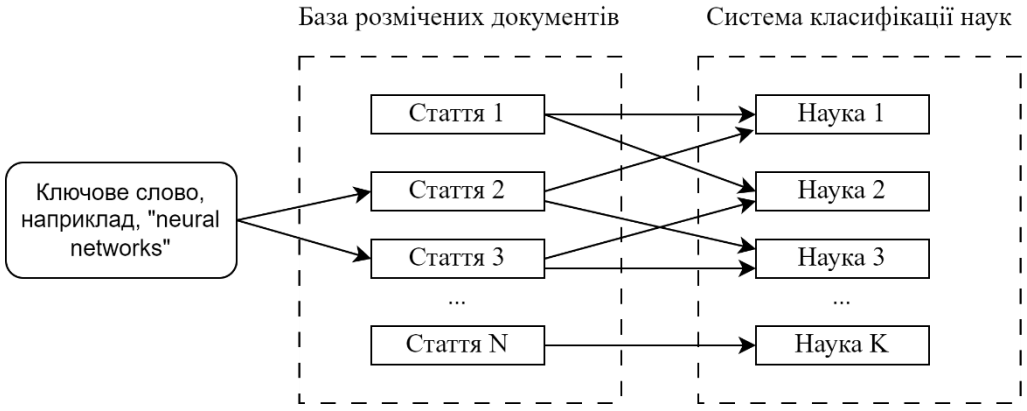

Рисунок 1. Схематичне зображення категоризації ключового слова

Внаслідок категоризації профіль заявки у формі множини її ключових слів $A_w = \{w_1, w_2, ..., w_n\}$ перетворюється у профіль заявки у формі категоріального розподілу за спеціальностям $A_t = \{\mu_{t_1}(A), \mu_{t_2}(A), ..., \mu_{t_m}(A)\}$, де $\mu_{t_i}(A) \in [0;1]$ – ступінь належності заявки A до спеціальності t_i, $i = \overline{1, m}$. Аналогічно, профіль рецензента у формі множини його ключових слів чи дослідницьких інтересів $R_w = \{w_1, w_2, ..., w_n\}$ перетворюється у профіль рецензента у формі категоріального розподілу за спеціальностям $R_t = \{\mu_{t_1}(R), \mu_{t_2}(R), ..., \mu_{t_m}(R)\}$.

Визначення схожості між профілями заявки та рецензента

Для зіставлення рецензентів та заявок потрібно знати, наскільки схожі 2 категоріальні розподіли – розподіл за спеціальностями ключових слів рецензента та розподіл за спеціальностями ключових слів заявки. У категоріальному просторі схожість двох об'єктів визначається зазвичай як суперпозиція схожості об'єктів за кожною категорією. Найчастіше – це сума схожості за окремими категоріями, коли кожна категорія розглядається

незалежно та ізольовано від інших. Усі метрики схожості з оглядових статей [23, 24] базуються на припущенні, що спорідненість між категоріями відсутня. Але деякі наукові спеціальності є спорідненими, зокрема для ANZSRC-2020 в [25] виявлено 20 пар сильно споріднених спеціальностей, 41 пару з середньою спорідненістю та 70 пар зі слабкою спорідненістю. Тому схожість доцільно розраховувати не лише напряму, як схожість між еквівалентними спеціальностями, але врахувати і перехресну схожість для споріднених спеціальностей. Для цього застосуємо метрику з [26]. Метрика розраховує схожість двох об'єктів X та Y із такими категоріальними розподілами $(\mu_1(X), \mu_2(X), ..., \mu_m(X))$ та $(\mu_1(Y), \mu_2(Y), ..., \mu_m(Y))$, де m – кількість категорій, якими в нашому випадку є наукові спеціальності, $\mu_i(X)$ – ступінь належності об'єкта X до i-ї категорії, $\mu_i(Y)$ – ступінь належності об'єкта Y до i-ї категорії, $i = \overline{1, m}$. Розподіли мають бути нормалізованими, тобто задовольняти такі умови:

$$\mu_i(X) \in [0;1], \qquad \mu_i(Y) \in [0;1], \qquad i = \overline{1, m};$$

$$\sum_{i=\overline{1, m}} \mu_i(X) = 1;$$

$$\sum_{i=\overline{1, m}} \mu_i(Y) = 1.$$

За метрикою [26], схожість об'єктів X та Y визначається так:

$$Fit(X, Y) = \sum_{i=\overline{1, m}} \min(\mu_i(X), \mu_i(Y)) + \Delta F(X, Y), \tag{1}$$

де $\sum_{i=\overline{1, m}} \min(\mu_i(X), \mu_i(Y))$ – доданок, що оцінює безпосередню (пряму) схожість об'єктів X та Y;

$\Delta F(X, Y)$ – доданок, що враховує схожість об'єктів X та Y через споріднені категорії.

Після розрахунку прямої схожості отримуємо такі залишки належності:

$r_i(X) = \max(0, \mu_i(X) - \mu_i(Y))$, $r_i(Y) = \max(0, \mu_i(Y) - \mu_i(X))$, $i = \overline{1, m}$.

Врахуємо внесок залишків у схожість двох об'єктів через спорідненість категорій. Вважатимемо, що інформація про попарну спорідненість категорій подана у формі такого бінарного відношення:

$$\mathbf{K} = \|k_{ij}\|,$$

де $k_{ij} \in [0;1]$– коефіцієнт спорідненості i-ї та j-ї категорій, $i = \overline{1,m}$, $j = \overline{1,m}$.

Чим більш подібні категорії, тим вищий коефіцієнт спорідненості. Відношення спорідненості є симетричним та рефлексивним, відповідно $k_{ij} = k_{ji}$ та $k_{ii} = 1$.

Композицію залишків представимо такою матрицею:

$$\mathbf{E} = \|e_{ij}\|,$$

де $e_{ij} = \min(r_i(X), r_j(Y))$, $i = \overline{1,m}$, $j = \overline{1,m}$.

Внесок залишків через перехресну спорідненість категорій розраховується так:

$$\Delta F(X,Y) = \sum_{i=1,\,m} \sum_{j=1,\,m} e_{ij} \cdot k_{ij} .$$

Підбір рецензентів як задача оптимізації

Розглядається задача підбору колективу рецензентів, які сукупно найкраще підходять для експертизи заявки. Для цієї задачі можливі 2 випадки: формування колективу з нуля та доповнення колективу новими членами.

Вважатимемо відомими профіль заявки $A_t = \{\mu_{t_1}(A), \mu_{t_2}(A), ..., \mu_{t_m}(A)\}$ та профілі k потенційних рецензентів $R_{tj} = \{\mu_{t_1}(R_j), \mu_{t_2}(R_j), ..., \mu_{t_m}(R_j)\}$, $j = \overline{1,k}$ у просторі з m спеціальностей. Усю множину рецензентів позначимо як $\mathbf{R} = \{R_1, R_2, .., R_k\}$.

Потрібно знайти підмножину рецензентів $S \subset \mathbf{R}$, яка має найбільшу сукупну відповідність заявці:

$$Fit(A, Agg(S)) \to \max .$$

де $Agg(S)$ - функція агрегації категоріальних розподілів множини відібраних рецензентів.

Агрегацію категоріальних розподілів за профілями рецензентів R_{tj}, $j = \overline{1,k}$ у просторі спеціальностей пропонується реалізувати за третім етапом алгоритму категоризації з [22].

Кількість рецензентів для однієї заявки позначимо через $c = |S|$. Ця кількість є сталою; зазвичай це від 2 до 5 осіб. Рівень відповідності між заявкою та колективом рецензентів розрахуємо за формулою (1).

Алгоритми підбору членів спецради

Задача підбору опонентів з математичного погляду – це пошук підмножини фіксованої потужності з деякої множини. Для вирішення таких задач на практиці застосовуються переважно наближені алгоритми. Серед множини можливих алгоритмів необхідно обрати той, який забезпечує баланс за показниками якості підбору та витратами ресурсів на знаходження розв'язку. У цій роботі пропонується використати такі алгоритми.

Повний перебір. Оптимальний варіант можна знайти повним перебором. Для заявки *A* необхідно серед усіх можливих *c*-ок із елементів множини потенційних рецензентів **R** обрати таку, що забезпечує максимальне значення відповідності. Складність такого перебору зростає майже експоненційно, тому навіть для задач середньої розмірності перебрати всі можливі варіанти та вкластися в якісь часові обмеження нереально. Причому кількість варіантів дуже сильно залежить від *c*. Наприклад, якщо потрібно обрати двох рецензентів зі 100, то потрібно перебрати 4 950 варіантів. А якщо якщо обирати трьох рецензентів зі 100, тоді кількість варіантів зросте до 161 700.

Повний перебір на проріджений множині кандидатів. На практиці кандидати з низьким рівнем схожості навряд чи будуть призначені рецензентами заявки. Тому раціональним кроком буде нехтування потенційними рецензентами з дуже низькою схожістю. Відкинувши

кандидатів із низькою схожістю з дисертацією, наприклад, на рівні 0.1 чи 0.2, можна відчутно скоротити перебір. Чим сильніше проріджуватимемо початковий список кандидатів, тим меншою буде тривалість оптимізації, але водночас зростають ризики задалеко відхилитися від оптимуму.

Жадібний алгоритм. Суть полягає у тому, що рецензенти підбираються ітеративно з забезпечення на кожному кроці максимальної відповідності поточного фрагмента колективу заявці. Алгоритм виконується за c ітерацій. На кожній ітерації до колективу рецензентів додається один новий член, який на цій ітерації максимізує рівень відповідності поточного складу заявці. На першій ітерації знаходимо кандидата з найвищою схожістю з дисертацією. На другій ітерації обираємо кандидата, який сукупно зі вже підібраним членом ради має найвищу відповідність дисертації. Кількість операцій перебору за такого підходу значно зменшується, але розв'язок може вийти неоптимальним. Для задачі підбору опонентів, окрім вище описаного класичного варіанта жадібного алгоритму, можливий варіант із елітизмом. Елітизм полягає в тому, що спочатку додається кандидат із найбільшим рівнем відповідності дисертації. При цьому рівень відповідності оновленого фрагмента ради дисертації не враховується. Далі, інші опоненти підбираються за класичним жадібним алгоритмом, тобто призначаються кандидати, які на поточній ітерації максимізують рівень відповідності колективу експертів дисертації. Жадібний алгоритм, особливо його елітарна реалізація, суттєво скорочують тривалість оптимізації.

Ізольований підбір. Найпростішим способом призначення рецензентів є обрання тих, які найбільш схожі з профілем дисертації. При цьому відповідність колективу рецензентів дисертації не враховується. Вважається, що чим сильніше кожен із кандидатів відповідає тематиці дисертації, тим кращою буде і разова рада. Грубо кажучи, вважається, що рівень відповідності ради є сумою рівнів відповідності кожного члена. Алгоритмічно, ізольований підбір реалізується сортуванням кандидатів за спаданням рівня схожості до дисертації та відбором перших c кандидатів. Це дуже швидкий алгоритм, але з малими шансами потрапити в оптимум. За такого алгоритму можливі ситуації, коли всі члени разової ради відповідатимуть лише одному і тому ж фрагмента тематики дисертації, а інші фрагменти тематики достовірно оцінити не буде кому.

Датасет дисертацій

Для експериментів із підбору рецензентів сформуємо датасет дисертацій [27]. Для цього скористаємось інформаційною системою НАЗЯВО – NAQA.Svr. У цю систему подаються заявки на разові спецради для їх затвердження Міністерством освіти і науки України. Кожна дисертація містить перелік супровідних документів та дані про разову раду, що пропонується закладом. Ми зібрали інформацію за 67 дисертаціями: 17 із них відхилені міністерством через слабку відповідність тематики статей членів ради дисертації, а 50 дисертацій було захищено. У ті 50 дисертацій увійшли і 17 раніше відхилених, для яких сформували нові ради. Зібрані дисертації належать до різних спеціальностей (рис. 2) з домінуванням спеціальностей 12-ї галузі, а саме 121, 122, 123, 124, 125 та 126. Кожен запис у датасеті включає унікальний ідентифікатор дисертації, прізвище та ім'я здобувача, ключові

слова дисертації англійською мовою, список членів разової ради разом з ключовими словами їх статей та статус ради – відхилена чи схвалена міністерством.

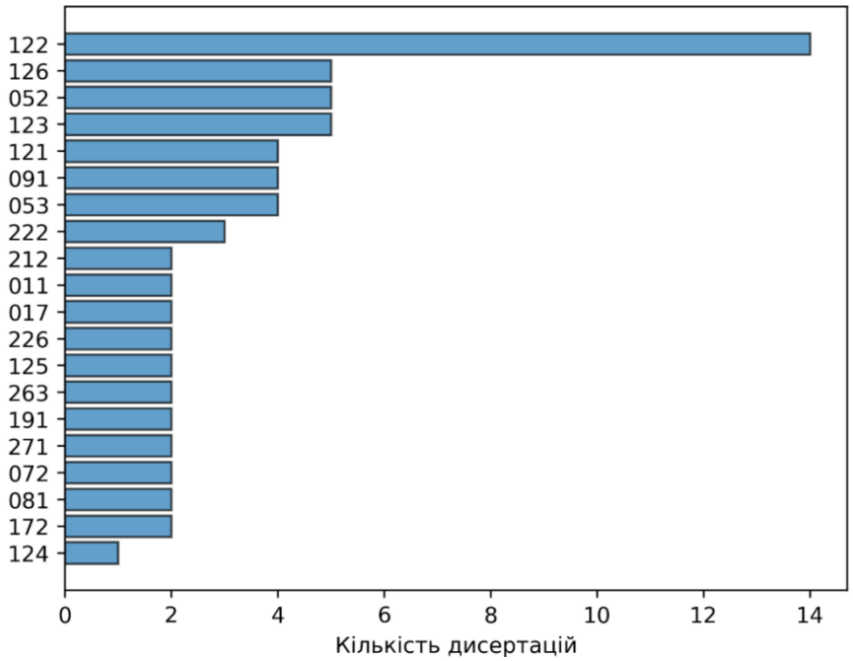

Рисунок 2. Розподіл дисертацій датасету за спеціальностями

На рис. 3 подано ранговий розподіл відповідності дисертації разовій раді, яка запропонована закладом. Більша частина разових рад мають відповідність вище 0.2 (62 ради), а решта мають незначущий рівень відповідності (5 рад). Міжквартильний інтервал приблизно дорівнює [0.4; 0.8].

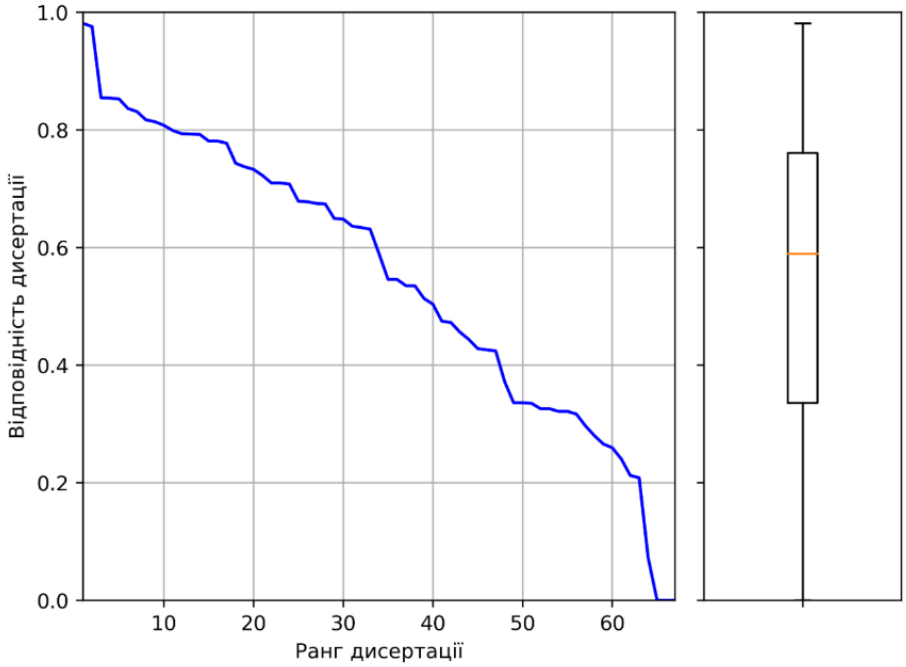

Рисунок 3. Ранговий розподіл відповідності дисертації разовій раді, яку сформував заклад

Експерименти з підбору опонентів

Проведемо експерименти з підбору опонентів на сформованому датасеті дисертацій. Для цього спочатку категоризуємо ключові слова дисертації за алгоритмом категоризації ключових слів у межах спеціальностей наук з ANZSRC-2020. Далі, аналогічним способом категоризуємо ключові слова статей членів разових рад. Пари ключових слів будемо поєднувати у додаткові запити лише в межах однієї статті. Для кожної ради вилучимо опонентів і спробуємо підібрати кращих із членів інших разових рад. Після вилучення опонентів отримуємо множину фрагментів разових рад, що містять голову та двох або одного рецензентів. Необхідно знайти опонентів, додавання яких до фрагментів разових рад забезпечить їх максимально можливу відповідність тематиці дисертацій.

Результати підбору опонентів порівняємо з варіантом разової ради, який сформовано закладом. Кількісно ефект оцінимо середнім рівнем зміни відповідності разових рад:

$$E = \frac{\sum\limits_{i=\overline{1,N}} F_i^{new} - F_i^{current}}{\sum\limits_{i=\overline{1,N}} F_i^{current}} \cdot 100\% ,$$

де N – кількість дисертацій;

F_i^{new} – рівень відповідності разової ради i-ї дисертації після оптимізації, $i = \overline{1,N}$;

$F_i^{current}$ – рівень відповідності разової ради i-ї дисертації до оптимізації, $i = \overline{1,N}$.

На рис. 4 подано результати оптимізації за різних алгоритмів підбору. Майже для всіх випадків разові ради від закладу мають нижчу відповідність тематиці дисертацій, ніж знайдені за будь-яким алгоритмом підбору. В умовах ручного формування рад закладом та обмежених можливостей щодо вибору членів рад отримуємо середній рівень відповідності тематиці. З іншого боку, за автоматичного підбору членів ради та достатнього великого пулу кандидатів отримуємо значне покращення рад лише за рахунок підбору опонентів.

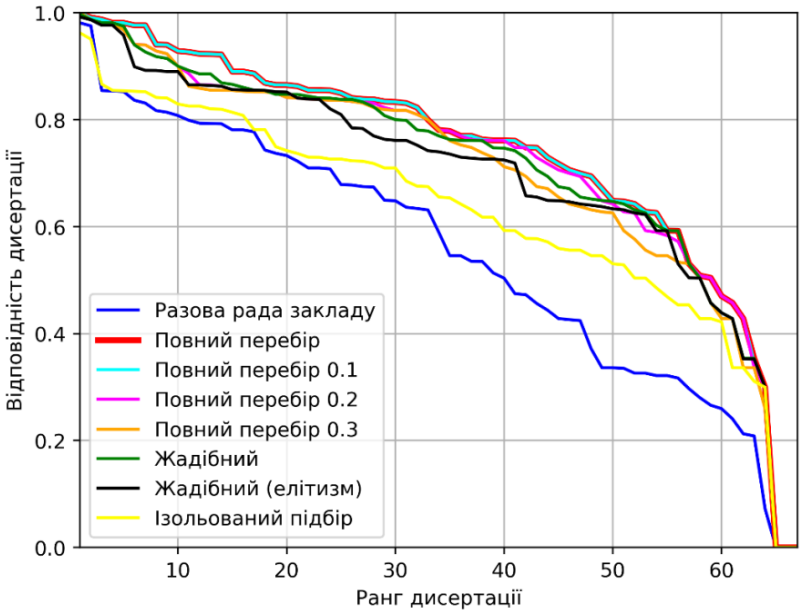

Рисунок 4. Результати підбору рад за різних алгоритмів оптимізації

На рис. 5 та 6 порівнюються рівні відповідності разових рад від закладу зі знайденими варіантами рад. За повного перебору наявне значне покращення відповідності більшості рад – як тих, які були відхилені, так і всіх інших. Деякі ради не покращено або рівень покращення низький. Це зумовлено насамперед тим, що в датасеті розподіл дисертацій за спеціальностями є нерівномірним (див. рис. 2) і датасет має малий обсяг. За найпростішого алгоритму на основі ізольованого підбору (див. рис. 6) більшість разових рад покращено, але відповідність кількох рад погіршилася. Погіршення пояснюється тим, що висока схожість кандидата з дисертацією не означає, що утворений за ізольованим підходом колектив буде покривати усю тематику дисертації.

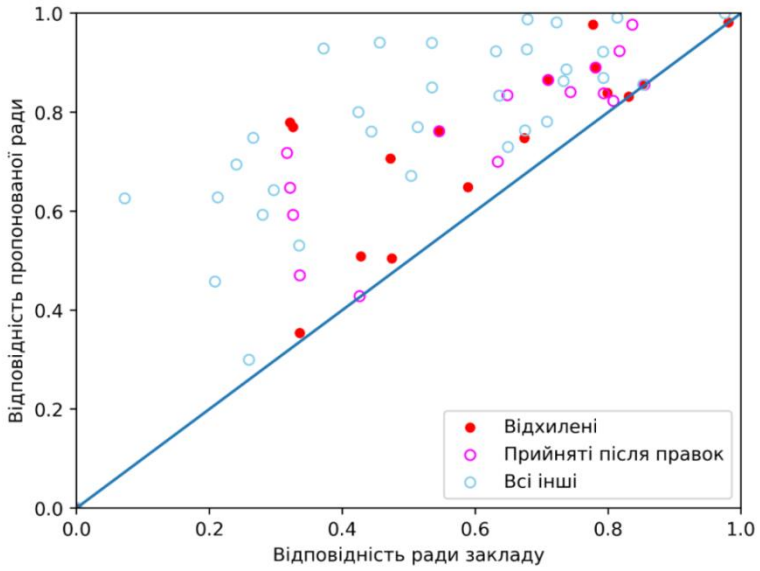

Рисунок 5. Порівняння рад від закладу з пропонованими радами, які сформовано за повним перебором

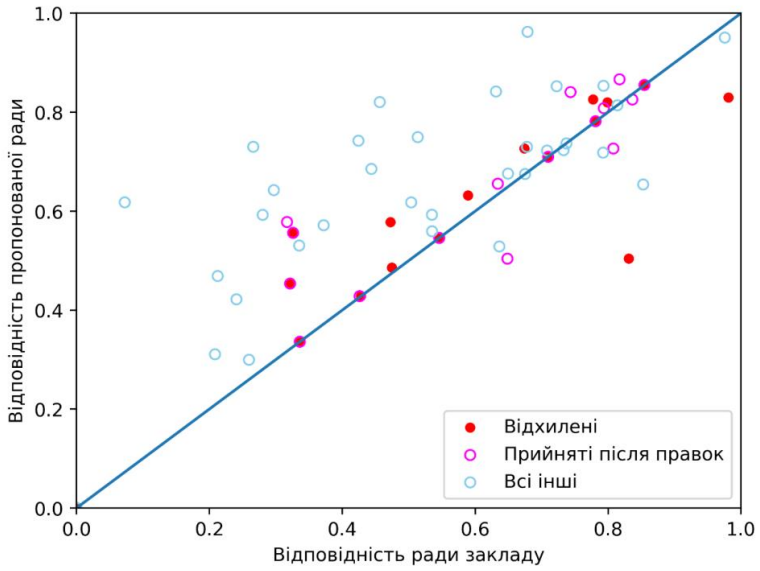

Рисунок 6. Порівняння рад від закладу з пропонованими радами, які сформовано за ізольованим підбором

На рис. 7 порівнюються результати підбору рад за різними алгоритмами оптимізації. Перебір на проріджений множині кандидатів за порогу у 0.3 однозначно невдалий. Усі інші утворюють множину Парето. Тому під час вибору алгоритму необхідно врахувати пріоритети – потрібен швидкий результат чи якісний. З рис. 7 видно, що рівень покращення внаслідок переходу з жадібного пошуку на алгоритми повного перебору зростає повільно. Але тривалість оптимізації зростає суттєво. Тому найбільш збалансованим можна вважати жадібний алгоритм підбору без елітизму. Альтернативою йому може бути алгоритм повного перебору з проріджуванням множини кандидатів за рівнем відповідності в околі 0.25. Але треба мати на увазі, що ці висновки ґрунтуються на експериментах на невеликому датасеті. За реальних баз даних великого обсягу тривалість оптимізації за алгоритмами повного перебору може зрости занадто сильно.

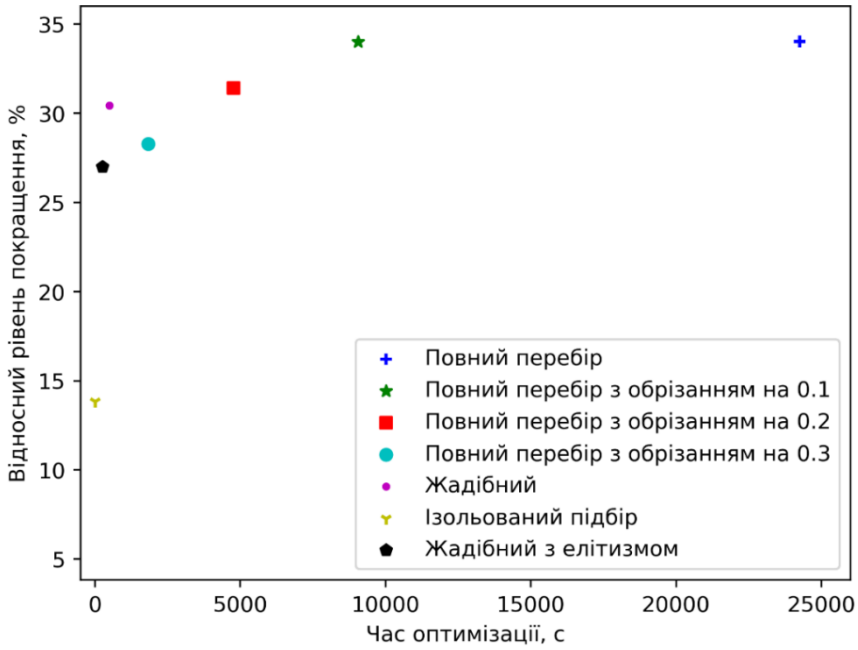

Рисунок 7. Оцінка алгоритмів підбору опонентів за критеріями «тривалість – якість»

Приклад покращення разової ради від закладу

Розглянемо, як внаслідок оптимізації вдалося покращити склад разової ради, порівняно зі складом від закладу, на прикладі дисертації *Моделі та методи обробки даних системи віддаленого моніторингу стану пацієнтів з цукровим діабетом* авторства *Віталія Левківського*; ідентифікатор дисертації в НАЗЯВО – *4756*.

Ключові слова дисертації:

edge devices; *IoT;*
diagnostics; *diseases;*
intelligent data analysis; *information technologies;*
medical information systems; *modeling;*
monitoring; *data processing;*
patient; *forecasting;*
software component model; *system design;*
diabetes.

З переліку ключових слів зрозуміло, що тематикою дисертації є медичні інформаційні системи і технології отримання та обробки медичних даних. Для категоризації початкову множину ключових слів розширимо попарним поєднанням ключових слів. Це дасть змогу під час категоризації надати перевагу тим науковим спеціальностям, у яких одночасно зустрічаються пари ключових слів [22].

Результат категоризації множини ключових слів дисертації є таким:

4605 Data Management and Data Science – 0.382;

4606 Distributed Computing and Systems Software – 0.255;

4609 Information Systems – 0.205;

4203 Health Services and Systems – 0.158.

Назві наукової спеціальності передує її чотирицифровий ідентифікатор в ANZSRC-2020. Саме ці ідентифікатори використовуватимемо для подальшого викладу матеріалу. Аналізована дисертація представляється таким вектором: $A_t = \left(\frac{0.382}{4605}, \frac{0.255}{4606}, \frac{0.205}{4609}, \frac{0.158}{4203} \right)$.

У базі НАЗЯВО тематику досліджень кожного члена ради представлено ключовими словами 3 або 4 його статей. Для їх категоризації застосуємо принцип мішка ключових слів. Категоризація відбувається так: 1) для кожної множини ключових слів однієї статті додаємо їх попарні поєднання; 2) об’єднуємо отримані множини ключових слів різних статей в одну множину – в один мішок; 3) категоризуємо отриману множину ключових слів за алгоритмом [22].

Ключові слова статей голови разової ради є такими:

1) *information technology, cardiovascular system, diagnostics, SCORE, systematic coronary risk evaluation, blood pressure, diseases, medical information systems;*

2) *information technology, data processing, data, data visualization, COVID-19, mobile application, disease, medical information systems;*

3) *differential equations, acceleration, cloud computing, task analysis, computational complexity, computer simulation, computerized integration procedure, integration procedure, numerical method.*

Результат категоризації голови ради вийшов таким:

4609 Information Systems – 0.381;

4203 Health Services and Systems – 0.225;

4606 Distributed Computing and Systems Software – 0.214;

4601 Applied Computing – 0.180.

Ключові слова статей першого рецензента є такими:

1) *IoT, monitoring system, microclimate parameters, educational institutions;*

2) *biotechnical system, edge device, photoplethysmograph, photoplethysmography, pulse wave, saturation sensor, cardiovascular system, data handling, edge computing, optoelectronic devices, quantum optics, circulatory disorders, functional state, mode of operations, non-contact sensors, photoplethysmographic, physical conditions, pulse wave signal, technical realization, digital storage;*

3) *pandemic, health-saving educational environment, model, information and digital environment;*

4) *planning technologies, use case, project, systems design, project complexity.*

Результат категоризації першого рецензента вийшов таким:

4606 Distributed Computing and Systems Software – 0.337;

4605 Data Management and Data Science – 0.256;

4003 Biomedical Engineering – 0.244;

3208 Medical Physiology – 0.162.

Ключові слова статей другого рецензента є такими:

1) *mathematical model, biotechnical system, edge devices, methods of studying, human body health, hemodynamic characteristics, diagnostic systems, digital signal processing, information technology;*

2) *biotechnical system, edge device, photoplethysmograph, photoplethysmography, pulse wavesaturationsensor, cardiovascular system, data handling, edge computing, optoelectronic devices, quantum optics, circulatory disorders, functional state, mode of operations, non-contact sensors, photoplethysmographic, physical conditions, pulse wave signal, technical realization, digital storage;*

3) *IoT, monitoring system, microclimate parameters, educational institutions, edge devices.*

Результат категоризації другого рецензента вийшов таким:

4606 Distributed Computing and Systems Software – 0.426;

4605 Data Management and Data Science – 0.299;

4003 Biomedical Engineering – 0.138;

4604 Cybersecurity and Privacy – 0.135.

Ключові слова статей першого опонента є такими:

1) *information expert system, control, method of fuzzy sets, medical diagnostics, diabetes, dentistry;*

2) *medical information technologies, medical information systems, coronary channels, coronary artery disease;*

3) *intelligent technologies, computer planning, modeling, medical diagnostics, nasal breathing testing, treatment, rehabilitation, medicine, tissue trophic complex, magnetic data analysis, electrocardiographic data analysis, audiological data analysis;*

4) *medical expert systems, fuzzy logic, coronary artery disease, coronary arteries, problems of cardiology, cardiovascular diseases, myocardial infarction, patient safety.*

Результат категоризації першого опонента вийшов таким:

3201 Cardiovascular Medicine and Haematology – 0.387;

3203 Dentistry – 0.215;

4605 Data Management and Data Science – 0.205;

4602 Artificial Intelligence – 0.192.

Ключові слова статей другого опонента є такими:

1) *Data analysis, Computational modeling, Neural networks, Predictive models, Safety, Personnel, Task analysis, hypertension, medical robotic platforms, TIMA, health and safety digital competencies;*

2) *Medical Diagnosis, Forecasting, Neuroevolution, Synthesis, Adaptive Mechanism, Genetic Algorithm, Parallel Genetic Algorithm, Crossover, Hospital data processing, Intelligent systems, Neural networks, Crossover operator, Medical diagnostics, Modern computer systems, Network synthesis, Neuro evolutions, Parallelizations, Resource consumption, Synthesis process, Diagnosis;*

3) *dermatoscopy, medical diagnosis, convolutional neural network, skin disease, ResNet50 model, software component model.*

Результат категоризації другого опонента вийшов таким:

4602 Artificial Intelligence – 0.435;

4611 Machine Learning – 0.357;

4605 Data Management and Data Science – 0.208.

Зведемо отримані результати в табл. 1. З неї видно, що всі члени разової ради мають значний рівень схожості з дисертацією. Агрегуємо результати категоризації усіх членів разової ради, щоб знайти рівень відповідності разової ради дисертації. Агрегування реалізуємо третім етапом алгоритму категоризації ключових слів з [22]. Внаслідок агрегування отримуємо:

$$Agg\left(\begin{array}{l}\left(\frac{0.381}{4609},\frac{0.225}{4203},\frac{0.214}{4606},\frac{0.180}{4601}\right)\\ \left(\frac{0.337}{4606},\frac{0.256}{4605},\frac{0.244}{4003},\frac{0.162}{3208}\right)\\ \left(\frac{0.426}{4606},\frac{0.299}{4605},\frac{0.138}{4003},\frac{0.135}{4604}\right)\\ \left(\frac{0.387}{3201},\frac{0.215}{3203},\frac{0.205}{4605},\frac{0.192}{4602}\right)\\ \left(\frac{0.435}{4602},\frac{0.357}{4611},\frac{0.208}{4605}\right)\end{array}\right)=\left(\frac{0.389}{4606},\frac{0.374}{4605},\frac{0.236}{4602}\right).$$

Таблиця 1. Результати категоризації профілів членів разової ради від закладу

Член разової ради	Профіль члена ради R_{jt}	Схожість між дисертацією та членом ради $Fit(A, R_j)$
Голова	$\left(\frac{0.381}{4609},\frac{0.225}{4203},\frac{0.214}{4606},\frac{0.180}{4601}\right)$	0.577
Рецензент 1	$\left(\frac{0.337}{4606},\frac{0.256}{4605},\frac{0.244}{4003},\frac{0.162}{3208}\right)$	0.564
Рецензент 2	$\left(\frac{0.426}{4606},\frac{0.299}{4605},\frac{0.138}{4003},\frac{0.135}{4604}\right)$	0.521
Опонент 1	$\left(\frac{0.387}{3201},\frac{0.215}{3203},\frac{0.205}{4605},\frac{0.192}{4602}\right)$	0.239
Опонент 2	$\left(\frac{0.435}{4602},\frac{0.357}{4611},\frac{0.208}{4605}\right)$	0.227

Рівень відповідності разової ради до дисертації становить: $Fit\left(\left(\frac{0.382}{4605},\frac{0.255}{4606},\frac{0.205}{4609},\frac{0.158}{4203}\right),\left(\frac{0.389}{4606},\frac{0.374}{4605},\frac{0.236}{4602}\right)\right)=0.631$. Це доволі непоганий рівень відповідності, який в основному обумовлено сильним співпадінням за двома спеціальностями із чотирьох, а саме за *4606 Distributed Computing and Systems Software* та *4605 Data Management and Data Science*. Під час розрахунку рівня відповідності також враховано спорідненості спеціальностей *4602* та *4605* та спеціальностей *4605* та *4606*, які дорівнюють 0.122 та 0.124, відповідно [25].

Спробуємо підібрати кращих опонентів, щоб підвищити сукупну відповідність разової ради дисертації. Як потенційних кандидатів візьмемо членів усіх інших разових рад сформованого датасету. Внаслідок оптимізації опонентами призначено *Оксану Мелеховець* із відповідністю тематиці дисертації на рівні 0.158 та *Камілу Сторчак* із відповідністю тематиці дисертації на рівні 0.542.

Ключові слова статей першого запропонованого опонента є такими:

1) *trophic ulcers, diabetes mellitus, chronic venous insufficiency, photodynamic therapy, plasma therapy;*

2) *diabetes mellitus, diabetic foot, photodynamic therapy, autologous plasma;*

3) *type 2 diabetes, middle age, obesity, quality of life, physical therapy program;*

4) *type 2 diabetes, middle age, obesity, physical therapy program.*

Результат категоризації першого запропонованого опонента вийшов таким:

3210 Nutrition and Dietetics – 0.274;

4203 Health Services and Systems – 0.261;

4202 Epidemiology – 0.251;

3205 Medical Biochemistry and Metabolomics – 0.214.

Ключові слова статей другого запропонованого опонента є такими:

1) *intelligent data analysis, data mining system, neural network;*

2) *web system, machine learning, web development, Heroku, Streamlit;*

3) *machine learning methods, neural network, smart home, failures, prediction, data analysis, decision theory, information technology.*

Результат категоризації другого запропонованого опонента вийшов таким:

4605 Data Management and Data Science – 0.555;

4611 Machine Learning – 0.306;

4609 Information Systems – 0.139.

Виконаємо агрегування профілів членів ради за нових опонентів:

$$Agg\left(\begin{array}{l}\left(\frac{0.281}{4611},\frac{0.269}{4605},\frac{0.228}{4602},\frac{0.222}{4608}\right)\\ \left(\frac{0.556}{4612},\frac{0.302}{4602},\frac{0.142}{4007}\right)\\ \left(\frac{0.457}{4611},\frac{0.196}{4603},\frac{0.183}{4605},\frac{0.163}{4609}\right)\\ \left(\frac{0.274}{3210},\frac{0.261}{4203},\frac{0.251}{4202},\frac{0.214}{3205}\right)\\ \left(\frac{0.555}{4605},\frac{0.306}{4611},\frac{0.139}{4609}\right)\end{array}\right)=\left(\frac{0.361}{4605},\frac{0.335}{4606},\frac{0.156}{4609},\frac{0.148}{4203}\right).$$

Рівень відповідності дисертації разовій раді з новими опонентами становить: $Fit\left(\left(\frac{0.382}{4605},\frac{0.255}{4606},\frac{0.205}{4609},\frac{0.158}{4203}\right),\left(\frac{0.361}{4605},\frac{0.335}{4606},\frac{0.156}{4609},\frac{0.148}{4203}\right)\right)=0.923$. Порівнюючи з разовою радою від закладу, бачимо суттєве покращення рівня відповідності – профіль нової ради описується всіма тими самими спеціальностями, що і дисертація. Покращення становить приблизно 46%.

З наведеного прикладу видно, що хоча індивідуальна схожість окремого члена ради дисертації може бути і посередньою, проте сукупно рівень відповідності ради може вийти високим. Це відбувається завдяки тому, що новий опонент покриває так звану мінорну частину тематики дисертації, яка знаходиться поза полем експертизи інших членів ради. В аналізованому прикладі це частина тематики, яка відповідає спеціальностям *4609 Information Systems* та *4203 Health Services and Systems* (рис. 8).

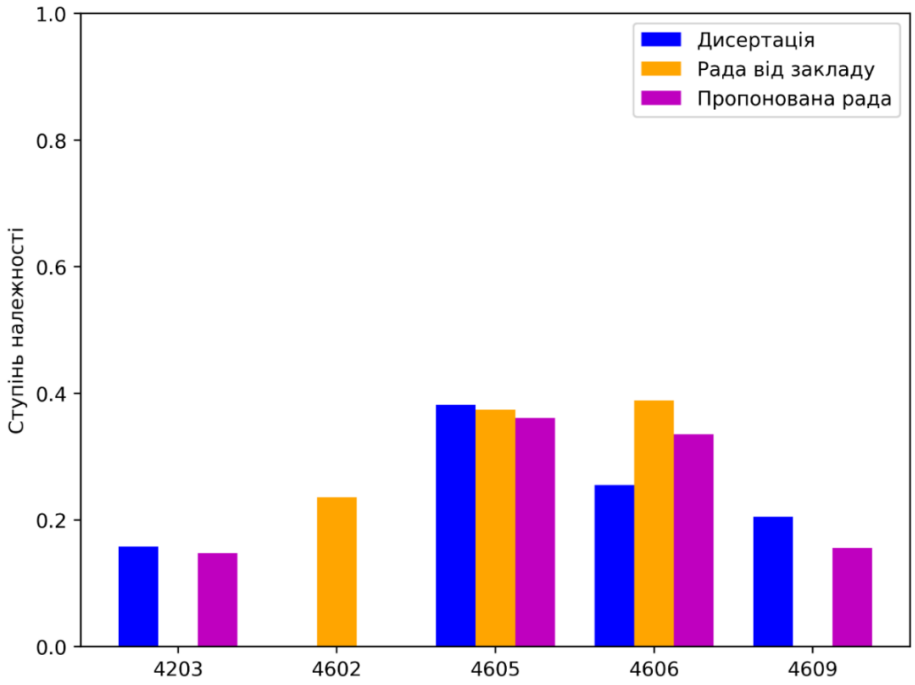

Рисунок 8. Порівняння категоризації дисертації та двох разових рад

Висновки

У роботі запропоновано швидкий метод підбору разової ради для захисту PhD-дисертацій, який, на відміну від ізольованого підбору рецензентів, враховує здатність саме ради як колективу спільно оцінити дисертацію за усіма аспектами її тематики. На першому етапі підбору здійснюється категоризація дисертації та потенційних членів разової ради шляхом представлення їх профілів як векторів у просторі наукових спеціальностей з ANZSRC-2020. На другому етапі розраховуються рівні відповідності потенційних членів ради тематиці дисертації з урахуванням спорідненості наукових спеціальностей. На третьому етапі підбирається склад ради, яка відповідає тематиці дисертації з максимально можливим ступенем. Для реалізації третього етапу запропоновано алгоритми на основі повного перебору на усій множині кандидатів, повного перебору на прорідженій множині кандидатів, на основі жадібного підходу без елітизму та з елітизмом, та за простим ізольованим пошуком. Тестування алгоритмів на сформованому датасеті із 67 PhD-дисертацій показало, що найкращий баланс за критеріями якості підбору та витрат ресурсів на пошук колективу забезпечує жадібний алгоритм без елітизму та повний перебір на прорідженій множині кандидатів. Внаслідок оптимізації вдалося покращити склад разових рад в середньому на 13–34% залежно від типу використаного алгоритму. Запропонований метод може

застосовуватись для покращення ефективності управління процесами призначення рецензентів для експертизи доповідей на конференціях, рецензування рукописів журнальних статей, експертизи наукових проєктів та грантових заявок, експертизи дисертацій тощо. Метод може використовуватися і в аудиторських цілях для швидкої перевірки коректності сформованих разових рад з подальшим відбором підозрілих справ для ґрунтовної ресурсовитратної експертизи. Подальші дослідження можуть стосуватися застосування запропонованого методу експрес-підбору рецензентів у більш трудомістких та ітеративних процедурах формування колективу рецензентів, коли потрібно враховувати не лише відповідність тематиці роботи, але й відсутність конфлікту інтересів, баланс навантаження на рецензентів та інші можливі обмеження. Подальші дослідження також варто зосередити на створенні більшого та репрезентативного датасету для ефективної оцінки якості різноманітних підходів до підбору рецензентів. Доцільно також під час формування ради враховувати не лише відповідність тематики рецензентів та дисертації, а також і кваліфікаційний рівень експертів.

Подяка

Автори висловлюють подяку Digital Science & Research Solutions Inc. за надання доступу до ресурсів Dimensions за проєктом DIM-371.

Література

1. Zhao, X., & Zhang, Y. (2022). Reviewer assignment algorithms for peer review automation: A survey. *Information Processing and Management*, *59*(5). https://doi.org/10.1016/j.ipm.2022.103028
2. Петричко, М. В., & Штовба, С. Д. (2024). Автоматизація підбору наукових рецензентів: огляд задач і методів. *Вісник Вінницького політехнічного інституту*, (1), 56-64. https://doi.org/10.31649/1997-9266-2024-172-1-56-64
3. Wang, F., Shi, N., & Chen, B. (2010). A comprehensive survey of the reviewer assignment problem. *International Journal of Information Technology and Decision Making*, *9*(4), 645-668. https://doi.org/10.1142/S0219622010003993
4. Aksoy, M., Yanik, S., & Amasyali, M. F. (2023). Reviewer assignment problem: A systematic review of the literature. *Journal of Artificial Intelligence Research*. AI Access Foundation. https://doi.org/10.1613/JAIR.1.14318
5. 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
6. Yarowsky, D., & Florian, R. (1999). Taking the load off the conference chairs: Towards a digital paper-routing assistant. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, EMNLP 1999* (pp. 220–230). Association for Computational Linguistics (ACL).
7. Karimzadehgan, M., Zhai, C. X., & Belford, G. (2008). Multi-aspect expertise matching for review assignment. In *Proceedings of International Conference on Information and Knowledge Management* (pp. 1113–1122). https://doi.org/10.1145/1458082.1458230

8. Mirzaei, M., Sander, J., & Stroulia, E. (2019). Multi-aspect review-team assignment using latent research areas. *Information Processing and Management*, *56*(3), 858–878. https://doi.org/10.1016/j.ipm.2019.01.007
9. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(4–5), 993–1022. https://doi.org/10.7551/mitpress/1120.003.0082
10. Ekinci, E., & Omurca, S. I. (2020). NET-LDA: A novel topic modeling method based on semantic document similarity. *Turkish Journal of Electrical Engineering and Computer Sciences*, *28*(4), 2244–2260. https://doi.org/10.3906/ELK-1912-62
11. Anjum, O., Gong, H., Bhat, S., Xiong, J., & Hwu, W. M. (2019). Pare: A paper-reviewer matching approach using a common topic space. In *EMNLP-IJCNLP 2019 – 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* (pp. 518–528). Association for Computational Linguistics. https://doi.org/10.18653/v1/d19-1049
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations ofwords and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. Neural information processing systems foundation.
13. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *EMNLP 2014 – 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (pp. 1532–1543). Association for Computational Linguistics (ACL). https://doi.org/10.3115/v1/d14-1162.
14. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051
15. Sun, C., Ng, K. T. J., Henville, P., & Marchant, R. (2019). Hierarchical word mover distance for collaboration recommender system. In *Communications in Computer and Information Science* (Vol. 996, pp. 289–302). Springer Verlag. https://doi.org/10.1007/978-981-13-6661-1_23
16. Kong, X., Jiang, H., Yang, Z., Xu, Z., Xia, F., & Tolba, A. (2016). Exploiting publication contents and collaboration networks for collaborator recommendation. *PLoS ONE*, *11*(2): e0148492. https://doi.org/10.1371/journal.pone.0148492
17. Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In *ACL 2018 – 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference* (Long Papers) (Vol. 1, pp. 328–339). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/p18-1031
18. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 – 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies – Proceedings of the Conference* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics (ACL).
19. Alec, R., Jeffrey, W., Rewon, C., David, L., Dario, A., & Ilya, S. (2019). Language models are unsupervised multitask learners | Enhanced Reader. *OpenAI Blog*, *1*(8), 9. Retrieved from https://github.com/codelucas/newspaper

20. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. ArXiv 2019. *arXiv preprint arXiv:1910.01108.*
21. Zhao, Y., Tang, J., & Du, Z. (2019). EFCNN: A restricted convolutional neural network for expert finding. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 11440 LNAI, pp. 96–107). Springer Verlag. https://doi.org/10.1007/978-3-030-16145-3_8
22. Shtovba, S., & Petrychko, M. (2021). An algorithm for topic modeling of researchers taking into account their interests in Google Scholar profiles. In *CEUR Workshop Proceedings* (Vol. 2864 "The Fourth International Workshop on Computer Modeling and Intelligent Systems", pp. 299–311). CEUR-WS. https://doi.org/10.32782/cmis/2864-26
23. Jie, Y., Amores, J., Sebe, N., & Qi, T. (2006). A new study on distance metrics as similarity measurement. In *2006 IEEE International Conference on Multimedia and Expo, ICME 2006 – Proceedings* (Vol. 2006, pp. 533–536). https://doi.org/10.1109/ICME.2006.262443
24. Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *City*, *1*(2), 1.
25. Штовба, С. Д., & Петричко, М. В. (2024). Ідентифікація рівня спорідненості наукових спеціальностей на основі даних системи Dimensions. *Проблеми програмування*, (1), 77–85. https://doi.org/10.15407/pp2024.01.077
26. Shtovba, S., Petrychko, M., & Shtovba, O. (2023). Similarity metric of categorical distributions for topic modeling problems with akin categories. In *CEUR Workshop Proceedings* (Vol. 3392 "The Sixth International Workshop on Computer Modeling and Intelligent Systems", pp. 76–85). CEUR-WS. https://doi.org/10.32782/cmis/3392-7
27. Petrychko, M., & Shtovba, S. (2024). Dataset for PhD theses reviewers assignments. *ResearchGate*. http://dx.doi.org/10.13140/RG.2.2.23147.35362

Рукопис отримано – 17/07/2024; прийнято до публікації – 29/07/2024.

Express assignment of reviewers for a PhD thesis defense committee

Serhiy Shtovba, Mykola Petrychko

Abstract

Today PhD thesis defense committee are formed manually. This causes both corruption risks and significant time spent on searching and analyzing candidates with a high chance of missing qualified opponents. Therefore, there is an interest in automating the formation of committees, which would allow to eliminate the mentioned risks of the human factor. The paper focuses on the express committee assignment when there is a need to narrow down a large list of candidates. The resulting short list can be analyzed either manually or processed by a fine-grained assignment procedure which is resource consuming and requires a much larger volume of initial information than the express assignment. A method of assigning a team of reviewers based on their relevance to the topic of the thesis is proposed, which, unlike the isolated assignment of candidates, takes into account the ability of the team of reviewers to jointly evaluate the work in terms of all aspects of its topic. The method is balanced in terms of assignment quality and resource costs criteria for the search of committee members. The method consists of 3 stages. At the first stage, the thesis and potential committee members are categorized by representing their topics with vectors in the space of research specialties from ANZSRC-2020. At the second stage, the level of correspondence of candidates to the topic of the thesis is calculated, taking into account the affinity of the research specialties of ANZSRC-2020. At the third stage, the committee is assigned, which corresponds to the topic of the thesis to the maximum possible extent. To implement the third stage, several optimization algorithms are proposed. Algorithm testing on the generated dataset of 67 PhD theses showed that the best balance in terms of assignment quality and resource costs criteria for team search provides a greedy algorithm without elitism and a complete search on a truncated set of candidates. As a result of the optimization, it was possible to improve the composition of committees by an average of 13-34%, depending on the type of algorithm used.

Keywords: reviewer assignment problem; express assignment; natural language processing; categorization; discrete optimization; data analysis; Dimensions.

References

1. Zhao, X., & Zhang, Y. (2022). Reviewer assignment algorithms for peer review automation: A survey. *Information Processing and Management, 59*(5). https://doi.org/10.1016/j.ipm.2022.103028
2. Petrychko, M., & Shtovba, S. (2024). Avtomatyzatsiia pidboru naukovykh retsenzentiv: Ohliad zadach i metodiv. *Visnyk Vinnytskoho politekhnichnoho instytutu*, (1), 56–64. https://doi.org/10.31649/1997-9266-2024-172-1-56-64
3. Wang, F., Shi, N., & Chen, B. (2010). A comprehensive survey of the reviewer assignment problem. *International Journal of Information Technology and Decision Making*, *9*(4), 645–668. https://doi.org/10.1142/S0219622010003993
4. Aksoy, M., Yanik, S., & Amasyali, M. F. (2023). Reviewer assignment problem: A systematic review of the literature. *Journal of Artificial Intelligence Research*. AI Access Foundation. https://doi.org/10.1613/JAIR.1.14318
5. 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
6. Yarowsky, D., & Florian, R. (1999). Taking the load off the conference chairs: Towards a digital paper-routing assistant. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, EMNLP 1999* (pp. 220–230). Association for Computational Linguistics (ACL).
7. Karimzadehgan, M., Zhai, C. X., & Belford, G. (2008). Multi-aspect expertise matching for review assignment. In *Proceedings of International Conference on Information and Knowledge Management* (pp. 1113–1122). https://doi.org/10.1145/1458082.1458230

8. Mirzaei, M., Sander, J., & Stroulia, E. (2019). Multi-aspect review-team assignment using latent research areas. *Information Processing and Management*, *56*(3), 858–878. https://doi.org/10.1016/j.ipm.2019.01.007
9. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(4-5), 993–1022. https://doi.org/10.7551/mitpress/1120.003.0082
10. Ekinci, E., & Omurca, S. I. (2020). NET-LDA: A novel topic modeling method based on semantic document similarity. *Turkish Journal of Electrical Engineering and Computer Sciences*, *28*(4), 2244–2260. https://doi.org/10.3906/ELK-1912-62
11. Anjum, O., Gong, H., Bhat, S., Xiong, J., & Hwu, W. M. (2019). Pare: A paper-reviewer matching approach using a common topic space. In *EMNLP-IJCNLP 2019 – 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* (pp. 518–528). Association for Computational Linguistics. https://doi.org/10.18653/v1/d19-1049
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations ofwords and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. Neural information processing systems foundation.
13. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *EMNLP 2014 – 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (pp. 1532–1543). Association for Computational Linguistics (ACL). https://doi.org/10.3115/v1/d14-1162.
14. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051
15. Sun, C., Ng, K. T. J., Henville, P., & Marchant, R. (2019). Hierarchical word mover distance for collaboration recommender system. In *Communications in Computer and Information Science* (Vol. 996, pp. 289–302). Springer Verlag. https://doi.org/10.1007/978-981-13-6661-1_23
16. Kong, X., Jiang, H., Yang, Z., Xu, Z., Xia, F., & Tolba, A. (2016). Exploiting publication contents and collaboration networks for collaborator recommendation. *PLoS ONE*, *11*(2): e0148492. https://doi.org/10.1371/journal.pone.0148492
17. Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In *ACL 2018 – 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference* (Long Papers) (Vol. 1, pp. 328–339). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/p18-1031
18. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 – 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies – Proceedings of the Conference* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics (ACL).
19. Alec, R., Jeffrey, W., Rewon, C., David, L., Dario, A., & Ilya, S. (2019). Language models are unsupervised multitask learners | Enhanced Reader. *OpenAI Blog*, *1*(8), 9. Retrieved from https://github.com/codelucas/newspaper
20. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. ArXiv 2019. *arXiv preprint arXiv:1910.01108.*
21. Zhao, Y., Tang, J., & Du, Z. (2019). EFCNN: A restricted convolutional neural network for expert finding. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 11440 LNAI, pp. 96–107). Springer Verlag. https://doi.org/10.1007/978-3-030-16145-3_8

22. Shtovba, S., & Petrychko, M. (2021). An algorithm for topic modeling of researchers taking into account their interests in Google Scholar profiles. In *CEUR Workshop Proceedings* (Vol. 2864 "The Fourth International Workshop on Computer Modeling and Intelligent Systems", pp. 299–311). CEUR-WS. https://doi.org/10.32782/cmis/2864-26
23. Jie, Y., Amores, J., Sebe, N., & Qi, T. (2006). A new study on distance metrics as similarity measurement. In *2006 IEEE International Conference on Multimedia and Expo, ICME 2006 - Proceedings* (Vol. 2006, pp. 533–536). https://doi.org/10.1109/ICME.2006.262443
24. Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *City*, *1*(2), 1.
25. Shtovba, S. & Petrychko, M., (2024). Identyfikatsiia rivnia sporidnenosti naukovykh spetsialnostei na osnovi danykh systemy Dimensions. *Problemy prohramuvannia*, (1), 77–85. https://doi.org/10.15407/pp2024.01.077
26. Shtovba, S., Petrychko, M., & Shtovba, O. (2023). Similarity metric of categorical distributions for topic modeling problems with akin categories. In *CEUR Workshop Proceedings* (Vol. 3392 "The Sixth International Workshop on Computer Modeling and Intelligent Systems", pp. 76–85). CEUR-WS. https://doi.org/10.32782/cmis/3392-7
27. Petrychko, M., & Shtovba, S. (2024). Dataset for PhD theses reviewers assignments. *ResearchGate*. http://dx.doi.org/10.13140/RG.2.2.23147.35362