

Міністерство освіти і науки України
Вінницький національний технічний університет

Кваліфікаційна наукова
праця на правах рукопису

ПЕТРИЧКО МИКОЛА ВОЛОДИМИРОВИЧ

УДК 004.852

ДИСЕРТАЦІЯ
ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ПІДБОРУ РЕЦЕНЗЕНТІВ ДЛЯ
ЕКСПЕРТИЗИ НАУКОВИХ ТВОРІВ

126 – Інформаційні системи та технології

12 – Інформаційні технології

Подається на здобуття наукового ступеня доктора філософії

Дисертація містить результати власних досліджень. Використання ідей,
результатів і текстів інших авторів мають посилання на відповідне джерело

_____ Микола ПЕТРИЧКО

Наукові керівники: Штовба Сергій Дмитрович,
доктор технічних наук, професор
Биков Микола Максимович,
кандидат технічних наук, доцент

Вінниця - 2024

АНОТАЦІЯ

Петричко М.В. Інформаційна технологія підбору рецензентів для експертизи наукових творів. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 126 «Інформаційні системи та технології». – Вінницький національний технічний університет Міністерства освіти і науки України, Вінниця, 2024.

У дисертаційній роботі розроблено інформаційну технологію підбору релевантних рецензентів на основі обмеженої інформації про рецензента, зокрема на основі його наукових інтересів, сформованих у вільному стилі. Це дозволить з усього списку рецензентів виділити тих, які відповідають напрямку рецензованого твору і яких доцільно відправити на процедуру детального аналізу відповідності тематики наукового твору.

У вступі наведено: актуальність теми дослідження, мету та завдання дослідження; об'єкт, предмет дослідження та наукова новизна отриманих результатів; зв'язок роботи з науковими програмами, планами, темами, грантами; практичну цінність отриманих результатів; апробацію матеріалів дисертації та публікації.

У першому розділі подано огляд стану проблеми та постановку задач дослідження. Проведено аналіз процесів експертизи наукових творів як об'єкту дослідження. Виділено ключові етапи вирішення задачі автоматичного підбору рецензентів для надання експертної оцінки. Проаналізовано дані про заявки та рецензентів, що використовуються для автоматичного підбору рецензентів. Здійснено огляд методів обробки природної мови в літературі, що використовуються для представлення текстових даних заявок та рецензентів. Розглянуто різні постановки задач оптимізації підбору рецензентів. Визначено задачі дослідження.

У другому розділі подано високорівнений опис алгоритму експрес-підбору рецензентів. Сформульовано загальну ідею методу представлення даних про заявки та рецензентів як розподіл належностей до спеціальностей науки в межах деякої системи класифікації наук. Для інформаційної бази дослідження обрано систему Dimensions, яка для категоризації публікацій використовує систему класифікації наук ANZSRC-2020.

У третьому розділі запропоновано критерій оцінювання схожості категоріальних розподілів, що враховує спорідненість категорій. Даний критерій використовується для визначення схожості між векторним представленням заявки та рецензента, де категоріями є наукові спеціальності. Проведено експериментальні дослідження запропонованого критерію та виявлено порогові рівні шумової спорідненості, які доцільно ігнорувати. Запропоновано використовувати індекс Жаккара для визначення спорідненості між двома науковими спеціальностями науки, що використовується у запропонованому критерії оцінювання схожості категоріальних розподілів. Здійснена ідентифікація рівня спорідненості наукових спеціальностей за системою ANZSRC-2020 та ідентифікація рівня спорідненості вітчизняних освітніх спеціальностей.

У четвертому розділі розроблено метод експрес-підбору рецензентів. Розроблено алгоритм категоризації ключових слів заявки та рецензентів в межах системи класифікації наук ANZSRC-2020. В результаті роботи алгоритму заявки та рецензенти представляються як своєрідні патерни в просторі наукових спеціальностей. Роботу алгоритму проілюстровано на покроковому контрольному прикладі. Сформовано навчальну вибірку даних про 2489 науковців і на її основі виконано автоматичний підбір параметрів запропонованого алгоритму. Наведено загальну постановку задачі експрес-підбору колективу рецензентів. Запропоновано та проаналізовано 7 алгоритмів призначення рецензентів, які забезпечують різні показники якості призначень та витрати ресурсів на пошук колективу рецензентів. Наведено приклад

визначення відповідності колективу рецензентів тематиці заявки в контексті задачі підбору рецензентів для рецензування PhD-дисертації.

У п'ятому розділі подано опис розроблених програмних компонентів інформаційної технології підбору рецензентів та загальну схему процесу підбору колективу рецензентів. Наведено контрольний приклад використання запропонованого алгоритму для підбору опонентів для рецензування дисертацій. Для контрольного прикладу сформовано датасет дисертацій та разових рад за відкритими даними Національного агентства із забезпечення якості вищої освіти. Проведено аналіз запропонованих алгоритмів призначень опонентів за показниками тривалості підбору та якістю разової ради. Показано детальний приклад формування колективу рецензентів на тестовій дисертації.

У висновках подані основні результати дисертаційного дослідження.

На основі досліджень, проведених у дисертаційній роботі отримано наступні наукові результати:

1. Удосконалено критерій Чекановського для оцінювання схожості категоріальних розподілів шляхом введення додаткового доданку, який враховує перехресну спорідненість категорій, що дозволяє взяти до уваги спорідненість наукових спеціальностей потенційного рецензента та наукового твору. Достовірність результату підтверджена обчислювальними експериментами, які показали, що на тестовій вибірці доданок який враховує спорідненість категорій має значну кількість викидів як для малої так і для великої кількості категорій.

2. Вперше розроблена модель статистичного оцінювання спорідненості наукових спеціальностей, яка ґрунтується на розрахунку індекса Жаккара за базою категоризованих наукових публікацій, що дозволяє швидко ідентифікувати рівень схожості спеціальностей в порівнянні з трудомісткими аналогами на основі аналізу списку цитованості та списку джерел кожної окремої статті. Достовірність результату підтверджена тестовими прикладами.

3. Вперше розроблено метод категоризації рецензентів, який базується на перетворенні обмеженого набору початкової інформації у формі наукових

інтересів рецензента або ключових слів заявки у вектор в просторі наукових спеціальностей за принципом word2vec, і відрізняється від формального співставлення шифрів спеціальностей аналізом фактичного змісту заявки та поточної компетентності рецензента, а від моделі співставлення ключових слів – врахуванням синонімії та спорідненості. Достовірність результату забезпечена налаштуванням параметрів алгоритму категоризації на тестовому датасеті з 2489 науковців та експериментально підтверджена для спеціальностей з ANZSRC-2020 з використанням інформаційної бази системи Dimensions.

4. Вперше розроблено метод призначення команди рецензентів за їх відповідністю тематиці заявки, який на відміну від ізольованого підбору рецензентів враховує здатність саме колективу рецензентів спільно оцінити заявку за усіма аспектами її тематики, та є збалансованим за критеріями якості підбору і витратами ресурсів на пошук колективу. Достовірність результату підтверджена шляхом проведення експериментів з підбору рецензентів на тестовому датасеті з 67 PhD-дисертацій.

Розроблена інформаційна технологія автоматизовує складні та трудомісткі процеси підбору рецензентів з використанням обмеженого набору початкової інформації. Окрім того, ідентифіковано рівні спорідненості наукових спеціальностей в межах Австралійсько-Новозеландської системи класифікації наук ANZSRC-2020 та рівні спорідненості вітчизняних науково-освітніх спеціальностей, що має окремо практичне значення для вирішення таких задач управління освітньо-науковою діяльністю, як корекція переліку спеціальностей та галузей знань, відслідковування трендів міждисциплінарних досліджень, формування перспективних міждисциплінарних освітніх програм та конкурсних пропозицій для широкого конкурсу, модернізації мережі факультетів та університетів. Також сформовано датасети для категоризації науковців та підбору рецензентів, що має окремо практичне значення для перевірки та удосконалення нових алгоритмів обробки такої інформації.

Практична цінність дисертації полягає в тому, що теоретичні складові інформаційної технології підбору рецензентів реалізовано програмно. Інформаційна технологія забезпечує ефективний експрес-підбір рецензентів з врахуванням здатності саме колективу рецензентів спільно оцінити заявку за усіма аспектами її тематики, та є збалансована за критеріями якості підбору і витратами ресурсів на пошук колективу. Практичну цінність підтверджує детальний приклад процесу підбору рецензентів в контексті задачі підбору опонентів до дисертацій. В результаті вдалося покращити склад разових рад, порівняно з радами від закладу, в середньому на 32-55%. Значення ефекту залежить від складності обраного алгоритму оптимізації.

Розроблена інформаційна технологія може використовуватись для покращення ефективності управління процесами підбору рецензентів в таких сферах як: рецензування на конференціях; підбір рецензентів для експертизи статей поданих у журнали; підбір рецензентів для експертизи наукових проєктів поданих для отримання грантів; підбір рецензентів та опонентів для дисертацій; аудит разових рад дисертацій – перевірка коректності сформованої ради відповідно до тематики дисертації. Обмеженням розробленої інформаційної технології є те, що вона не вирішує усі проблеми, які мають місце під час автоматичного підбору рецензентів, наприклад, у предметну область дослідження не входять питання усунення конфлікту інтересів, завантаженість рецензентів, формування актуального реєстру потенційних рецензентів тощо.

Ключові слова: аналіз даних, датасет, дискретна оптимізація, задача про призначення рецензентів, інформаційна технологія, категоризація, категоріальний розподіл, наукометрія, обробка природної мови, патерн наукової публікації, прийняття рішень, рекомендації, система класифікації наук, спорідненість наукових спеціальностей, тематика, Dimensions.

ABSTRACT

Petrychko M.V. Information technology for reviewer assignment problem in the field of research expertise. – Qualifying scientific work with manuscript rights.

The thesis is for obtaining the scientific degree of Doctor of Philosophy in specialty 126 «Information systems and technologies». – Vinnytsia National Technical University, Vinnytsia, 2024.

In the thesis an information technology is developed for assigning relevant reviewers based on limited information about a reviewer, in particular based on his/her scientific interests, formed in a free style. This will make it possible to assign from the entire list of reviewers those who correspond to the direction of the reviewed work and who should be sent to the procedure for a detailed analysis of reviewers relevance to the scientific work's topic.

The introduction states: the relevance of the research subject, the purpose and task of the research; object, scope of the research and scientific novelty of the obtained results; connection of work with scientific programs, plans, topics, grants; practical value of the obtained results; approval of thesis and publication materials.

The first chapter provides an overview of the state of the problem and setting research objectives. The analysis of scientific works' reviewing processes as an object of research was carried out. The key stages of solving the problem of automatic assignment of reviewers for providing expert evaluation are highlighted. Data on applications and reviewers used for automatic assignment of reviewers were analyzed. A review of natural language processing methods in the literature, used to represent textual data of applications and reviewers, was carried out. Various problem statements for optimizing the assignment of reviewers are considered. Research objectives are defined.

In the second chapter, a high-level description of the reviewers fast-assignment algorithm is presented. The general idea of the method of representing data on applications and reviewers as a distribution of memberships to research specialties within some research classification system is formulated. The Dimensions system

was chosen for the research information base, which uses the Australian and New Zealand Standard Research Classification (ANZSRC-2020) research classification system to categorize publications.

In the third chapter, a criterion for evaluating the similarity of categorical distributions is proposed, which takes into account the kinship of categories. This criterion is used to determine the similarity between the vector representation of an application and a reviewer, where the categories are research specialties. Experimental studies of the proposed criterion were conducted and threshold levels of noise kinship that should be ignored were revealed. It is proposed to use Jaccard index to determine the kinship between two research specialties, which is used in the proposed criterion for evaluating the similarity of categorical distributions. Identification of research specialties' kinship according to ANZSRC-2020 system and identification of Ukrainian educational specialties' kinship was carried out.

In the fourth chapter, the method of reviewers fast-assignment is developed. An algorithm for categorizing application keywords and reviewers within the framework of the ANZSRC-2020 research classification system has been developed. As the result, the algorithm represents applications and reviewers as patterns in the space of scientific specialties. The operation of the algorithm is illustrated by a step-by-step control example. A training sample of data on 2489 researchers was formed, and based on it, the parameters of the proposed algorithm were automatically selected. The general formulation of fast-assignment of a team of reviewers problem is given. 7 algorithms for the assignment of reviewers are proposed and analyzed, which provide various indicators of the quality of assignments and the expenditure of resources for the search of a team of reviewers. An example of determining the appropriateness of a team of reviewers to the topic of the application in the context of the task of assigning reviewers for reviewing a PhD thesis is given.

The fifth chapter provides a description of the developed software components of information technology for the assignment of reviewers and the general schema of assigning a team of reviewers to a proposal. A control example of the use of the proposed algorithm for the assignment of opponents for reviewing theses is given.

For a control example, a dataset of theses and one-time council was created based on open data of the National Agency for Quality Assurance of Higher Education. An analysis of the proposed algorithms for assigning of opponents was carried out according to indicators of the assignment time and the quality of the one-time council. A detailed example of assigning a team of reviewers to a test thesis is shown.

The main results of the thesis research are presented in the conclusions.

Based on the research conducted in the thesis, the following scientific results were obtained:

1. The Czekanowski criteria for evaluating the similarity of categorical distributions has been improved by introducing an additional term that takes into account the cross-similarity of categories, which allows taking into account the kinship of the scientific specialties of a potential reviewer and a scientific work. The reliability of the result was confirmed by computational experiments, which showed that on the test sample the proposed criterion is sensitive to the influence of related categories even for large-dimensional distributions.

2. For the first time, a model of statistical assessment of scientific specialties' similarity was developed, it is based on the calculation of the Jaccard index on the base of categorized scientific publications, which allows to quickly identify the level of similarity of specialties in comparison with time-consuming analogues that are based on the analysis of the list of citations and the list of sources of each individual article. The reliability of the result is confirmed by test examples.

3. For the first time, a reviewer categorization method was developed, which is based on the transformation of a limited set of initial information in the form of the reviewer's scientific interests or application keywords into a vector in the space of scientific specialties according to the word2vec principle, and differs from the formal comparison of specialty codes by the analysis of the actual content of the application and the reviewer's current competence, and from the keyword comparison model – by taking into account synonymy and relatedness. The reliability of the result is ensured by adjusting the parameters of the categorization algorithm on a test dataset

of 2,489 scientists and experimentally confirmed for specialties from ANZSRC-2020 using the Dimensions system information base.

4. For the first time, the method of assigning a team of reviewers based on their relevance to the topic of the proposal was developed, which in contrast to the isolated assignment of reviewers, it takes into account the ability of the team of reviewers to jointly evaluate the application on all aspects of its topic, and is balanced according to the criteria of the quality of the assignment and the cost of resources for the search of the team. The reliability of the result was confirmed by conducting experiments of reviewers assignment on a test dataset of 67 PhD theses.

The developed information technology automates the complex and time-consuming process of assigning reviewers using a limited set of initial information. In addition, the levels of kinship of scientific specialties within the Australian-New Zealand ANZSRC-2020 research classification system and the level of kinship of Ukrainian educational specialties have been identified, which has a separate practical significance for solving such problems of managing educational and scientific activities as correcting the list of specialties and fields of knowledge, tracking trends in interdisciplinary research, forming promising interdisciplinary educational programs and competitive proposals for a wide competition, modernization of the network of faculties and universities. Datasets for the categorization of scientists and the assignment of reviewers have also been created, which has a separate practical value for testing and improving new algorithms for processing such information.

The practical value of the thesis lies in the fact that the theoretical components of the information technology for the assigning of reviewers are implemented programmatically. Information technology provides effective fast assigning of reviewers, taking into account the ability of the team of reviewers to jointly evaluate the application on all aspects of its subject matter, and is balanced according to the criteria of assignment quality and resource costs for the search of the team. The practical value is confirmed by a detailed example of the process of selecting reviewers in the context of the task of selecting opponents for theses. As a result, it was possible to improve the composition of one-time councils, compared to councils

from the institution, by an average of 32-55%. The value of the effect depends on the complexity of the chosen optimization algorithm.

The developed information technology can be used to improve the efficiency of managing reviewer assignment processes in such areas as: reviewing at conferences; assignment of reviewers for examination of articles submitted to journals; assignment of reviewers for examination of scientific projects submitted for grants; assignment of reviewers and opponents for theses; audit of one-time dissertation councils – checking the correctness of the formed one-time council in accordance with the subject of the thesis. The limitation of the developed information technology is that it does not solve all the problems that occur during the automatic selection of reviewers, for example, the subject area of the study does not include issues of conflict of interest elimination, workload of reviewers, formation of an up-to-date register of potential reviewers, etc.

Keywords: data analysis, dataset, discrete optimization, reviewer assignment problem, information technology, categorization, categorical distribution, scientometrics, natural language processing (NLP), pattern of scientific publication, decision making, recommendations, research classification system, kinship of research specialties, subject matter, Dimensions.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Наукові праці, в яких опубліковані основні наукові результати дисертації:

– матеріали дисертаційної роботи, викладені у виданнях, що внесені до переліку фахових для захисту дисертацій з технічних наук:

[1] Shtovba, S., & Petrychko, M. (2019). Jaccard index-based assessing the similarity of research fields in dimensions. In CEUR Workshop Proceedings (Vol. 2533, pp. 117–128). CEUR-WS. (Входить до Scopus)

(Особистий внесок: Розробка метрики схожості між двома науковими спеціальностями ANZSRC, запропоновано використовувати індекс Жаккара, проведено аналіз даних за різні періоди)

[2] Shtovba, S., & Petrychko, M. (2021). An informetric assessment of various research fields interactions on base of categorized papers in dimensions. In CEUR Workshop Proceedings (Vol. 2845, pp. 159–169). CEUR-WS. (Входить до Scopus)

(Особистий внесок: Розробка метрики схожості між двома науковими спеціальностями ANZSRC, проведено аналіз даних за різні періоди та з різних галузей, виявлено тенденції зміни спорідненості)

[3] 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, pp. 299–311). CEUR-WS. <https://doi.org/10.32782/cmis/2864-26>. (Входить до Scopus)

(Особистий внесок: Розробка алгоритму категоризації ключових слів в межах деякої системи класифікації наук)

[4] Штовба С.Д., Петричко М.В. (2021). Тематичне моделювання науковців на основі їх інтересів у Google Scholar, Системні дослідження та інформаційні технології, №2, вересень, с. 113-129. ISSN: 1681-6048. <https://doi.org/10.20535/SRIT.2308-8893.2021.2.09>. (Входить до Scopus)

(Особистий внесок: Розробка алгоритму категоризації ключових слів в межах деякої системи класифікації наук)

[5] Штовба С.Д., Петричко М.В., Петранова М.Ю. (2023). Метрика схожості категоріальних розподілів, що враховує спорідненість різних категорій, Вісник Вінницького політехнічного інституту, №2, с. 49-57. <https://doi.org/10.31649/1997-9266-2023-167-2-49-57>. (Фахове видання)

(Особистий внесок: Розробка метрики порівняння категоріальних розподілів враховуючи перехресну спорідненість категорій)

[6] 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, pp. 76–85). CEUR-WS. <https://doi.org/10.32782/cmis/3392-7>. (Входить до Scopus).

(Особистий внесок: Розробка метрики порівняння категоріальних розподілів враховуючи перехресну спорідненість категорій)

[7] Штовба С. Д., Петричко М. В. (2024). Ідентифікація рівня спорідненості освітніх спеціальностей на основі аналізу профілей експертів НАЗЯВО. Наукові праці ВНТУ, №1. <https://doi.org/10.31649/2307-5376-2024-1-48-59>. (Фахове видання)

(Особистий внесок: Сформовано датасет та адаптовано метрику на основі індекса Жаккара для ідентифікації спорідненості освітньо-наукових спеціальностей)

[8] Штовба С. Д., Петричко М. В. (2024). Ідентифікація рівня спорідненості наукових спеціальностей на основі даних системи Dimensions. Проблеми програмування, №1. <https://doi.org/10.15407/pp2024.01.077>. (Фахове видання)

(Особистий внесок: Розроблено програмне забезпечення і ідентифіковано рівні спорідненості наукових спеціальностей ANZSRC-2020)

[9] Петричко М. В., Штовба С. Д. (2024). Автоматизація підбору наукових рецензентів: огляд задач і методів. Вісник Вінницького політехнічного інституту, №1, с. 56–64. <https://doi.org/10.31649/1997-9266-2024-172-1-56-64>. (Фахове видання)

(Особистий внесок: Проведено огляд літератури щодо автоматизації задач підбору рецензентів. Наведено етапи розв'язання задачі автоматичного підбору рецензентів та окреслено перспективні напрямки подальших досліджень)

Документи інтелектуальної власності, що отримані за результатами дисертації:

Штовба С.Д., Петричко М.В., Баєв А.В. Авторське свідоцтво на комп'ютерну програму «Double SimMetrics – метрика схожості категоріальних розподілів за подібних категорій», №116605 від 01.03.2023 р.

Наукові праці, які засвідчують апробацію матеріалів дисертації:

[1] Петричко М. В., Штовба С. Д. (2020). Автоматична категоризація науковців за їх інтересами в Google Scholar, на XV міжнародній конференції «Контроль і управління в складних системах», Вінниця.

(Особистий внесок: Розробка алгоритму категоризації ключових слів в межах деякої системи класифікації наук)

[2] Petrychko, M., Shtovba S. (2020). An Informetric Assessment of Various Research Fields Interactions on Base of Categorized Papers in Dimensions, in IT&I-2020: VII International conference «Information Technology and Interactions» – Satellite conference, Kyiv, pp. 170-171.

(Особистий внесок: Розробка метрики схожості між двома науковими спеціальностями, запропоновано використовувати індекс Жаккара, проведено аналіз даних за різні періоди)

[3] Петричко М. В., Штовба С. Д. (2021). Статистична оцінка рівня міждисциплінарності наукових досліджень з використанням даних системи Dimensions, на «Матеріали наукової конференції професорсько-викладацького складу, наукових працівників і здобувачів наукового ступеня за підсумками науково-дослідної роботи за період 2019–2020 рр.», Донецький національний університет імені Василя Стуса.

(Особистий внесок: Розробка метрики схожості між двома науковими спеціальностями, запропоновано використовувати індекс Жаккара, проведено аналіз даних за різні періоди)

[4] Петричко М. В., Штовба С. Д. (2021). Ідентифікація наукових спеціальностей дослідників на основі їх інтересів в Google Scholar, на Матеріали II всеукраїнської науково-практичної конференції студентів, аспірантів та молодих вчених «Прикладні інформаційні технології», Вінниця, Донецький національний університет імені Василя Стуса, с.70-73.

(Особистий внесок: Розробка алгоритму категоризації ключових слів в межах деякої системи класифікації наук)

[5] Петричко М. В., Штовба С. Д., Биков М.М. (2022). Налаштування алгоритму категоризації дослідників з використанням інформаційних ресурсів системи Dimensions, на XVI міжнародній конференції «Контроль і управління в складних системах», Вінниця.

(Особистий внесок: Оптимізація підбору параметрів алгоритму категоризації)

[6] Петричко М. В., Штовба С. Д. (2023). Метрика схожості категоріальних розподілів зі спорідненими категоріями, на VIII міжнародній науково-технічній конференції «Комп'ютерне моделювання та оптимізації складних систем», Дніпро.

(Особистий внесок: Розробка метрики порівняння категоріальних розподілів враховуючи перехресну спорідненість категорій)

[7] Петричко М. В., Штовба С. Д. (2024). Виявлення трендів мультидисциплінарних досліджень у 2022–2023 роках з використанням системи Dimensions. Прикладні аспекти сучасних міждисциплінарних досліджень, 200-202, Вінниця.

(Особистий внесок: Розробка програмного забезпечення та аналіз даних для виявлення трендів міждисциплінарних досліджень на основі моделі спорідненості наукових спеціальностей)

[8] Петричко М. В., Штовба С. Д. (2021). Модифікована метрика Чекановського для оцінки схожості науковців з врахуванням спорідненості спеціальностей, на Матеріали конференції «L Науково-технічна конференція підрозділів Вінницького національного технічного університету», Вінниця.

(Особистий внесок: Розробка метрики порівняння категоріальних розподілів враховуючи перехресну спорідненість категорій)

[9] Петричко М. В., Штовба С. Д., Тилець Р. О. (2020). Автоматична категоризація науковців на основі профілей в Google Scholar з врахуванням спорідненості між науковими спеціальностями, на Матеріали конференції «XLIX Науково-технічна конференція підрозділів Вінницького національного технічного університету», Вінниця.

(Особистий внесок: Розробка алгоритму категоризації ключових слів в межах деякої системи класифікації наук)

ЗМІСТ

ВСТУП.....	20
РОЗДІЛ 1 АНАЛІЗ СТАНУ ПРОБЛЕМИ ТА ПОСТАНОВКА ЗАДАЧ ДОСЛІДЖЕННЯ.....	27
1.1 Аналіз процесів експертизи наукових творів як об’єкту дослідження	27
1.2 Задача підбору рецензента для надання експертної оцінки	30
1.3 Аналіз даних, що використовуються для побудови профілів рецензента та заявки	35
1.4 Аналіз методів обробки даних про рецензентів та заявки	37
1.4.1 Сумка слів	38
1.4.2 TF-IDF	39
1.4.3 Word2Vec	40
1.4.4 Glove	41
1.4.5 Тематичне моделювання	43
1.5 Аналіз методів підбору рецензентів на основі інформаційного пошуку.....	43
1.5.1 Ручні підходи.....	43
1.5.2 Підходи на основі аналізу тексту/змісту	45
1.5.3. Підходи на основі нейронних мереж	50
1.6 Аналіз методів для підбору рецензентів на основі оптимізації	55
1.7 Дослідницька ідея та задачі подальшого дослідження	58
Висновки до розділу 1.....	60
РОЗДІЛ 2 УЗАГАЛЬНЕНИЙ ОПИС АЛГОРИТМУ ПІДБОРУ РЕЦЕНЗЕНТА.....	62
2.1 Загальна постановка задачі експрес-підбору рецензента	62
2.2 Представлення даних заявки та рецензента	63
2.2.1 Високорівневий опис алгоритму категоризації ключових слів в межах системи класифікації наук	64
2.2.2 Обґрунтування вибору системи класифікації наук та бази наукових публікацій	65
Висновки до розділу 2.....	70
РОЗДІЛ 3 СХОЖІСТЬ РЕЦЕНЗЕНТА ТА ТВОРУ З УРАХУВАННЯМ СПОРІДНЕНOSTІ НАУКОВИХ СПЕЦІАЛЬНОСТЕЙ.....	72

	18
3.1 Критерій схожості категоріальних розподілів з врахуванням схожості категорій	72
3.2 Експериментальні дослідження запропонованого критерію	77
3.3 Визначення рівня спорідненості наукових спеціальностей	82
3.4 Оцінка рівня спорідненості вітчизняних спеціальностей	90
Висновки до розділу 3.....	95
РОЗДІЛ 4 РОЗРОБКА МЕТОДУ ЕКСПРЕС-ПІДБОРУ РЕЦЕНЗЕНТІВ...	97
4.1 Формалізована постановка задачі категоризації ключових слів в межах системи класифікації наук.....	97
4.2 Категоризація ключових слів до спеціальностей науки з ANZSRC-2020	98
4.3 Контрольний приклад категоризації ключових слів	102
4.4 Підбір параметрів алгоритму категоризації.....	105
4.4.1 Ідея підбору параметрів алгоритму категоризації.....	105
4.4.2 Приклад використання запропонованого критерію схожості для порівняння розподілів на основі статей та на основі ключових слів	106
4.4.3 Формування датасету	108
4.4.4 Підбір параметрів.....	109
4.5 Категоризація за мішком ключових слів	115
4.6 Підбір колективу рецензентів.....	116
4.6.1 Загальна постановка задачі підбору рецензентів	116
4.6.2 Алгоритми вирішення задачі підбору рецензентів	117
4.6.2.1 Повний перебір.....	117
4.6.2.2 Повний перебір на обрізаній множині кандидатів	117
4.6.2.3 Жадібний алгоритм.....	118
4.6.2.4 Ізольований підбір	118
4.6.3 Оцінювання відповідності колективу рецензентів до заявки	119
Висновки до розділу 4.....	122
РОЗДІЛ 5 ПРОГРАМНА РЕАЛІЗАЦІЯ ТА ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ	124
5.1 Програмне забезпечення інформаційної технології.....	124
5.1.1 Архітектура системи.....	124
5.1.2 Програмний компонент парсингу даних щодо розподілів ключових слів	125

	19
5.1.3 Програмний компонент категоризації ключових слів	126
5.1.4 Програмний компонент критерію оцінювання схожості категоріальних розподілів	128
5.1.5 Програмний компонент формування тестових даних для оптимізації параметрів алгоритму категоризації ключових слів	128
5.1.6 Програмний компонент оптимізації параметрів алгоритму категоризації ключових слів	129
5.1.7 Програмний компонент оптимізації призначень опонентів до дисертацій	131
5.1.8 Загальна схема процесу підбору колективу рецензентів	131
5.2 Формування датасету дисертацій	132
5.3 Підбір опонентів	136
5.3.1 Оцінка ефекту від оптимізації	138
5.3.2 Оптимізація за повним перебором	139
5.3.3 Оптимізація за жадібним алгоритмом	142
5.3.4 Оптимізація за жадібним алгоритмом з елітизмом	143
5.3.5 Оптимізація за найбільш схожими кандидатами	145
5.3.6 Аналіз результатів оптимізації	146
5.3.7 Приклад покращення разової ради від закладу	147
Висновки до розділу 5.....	154
ВИСНОВКИ	156
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ	159
ДОДАТКИ.....	181
ДОДАТОК А. ПЕРЕЛІК ПУБЛІКАЦІЙ ЗДОБУВАЧА	182
ДОДАТОК Б. РЕЗУЛЬТАТИ ІДЕНТИФІКАЦІЇ СПОРІДНЕНOSTІ СПЕЦІАЛЬНОСТЕЙ З ANZSRC-2020	185
ДОДАТОК В. ПЕРЕЛІК ВІТЧИЗНЯНИХ ОСВІТНІХ СПЕЦІАЛЬНОСТЕЙ ..	195
ДОДАТОК Г. РЕЗУЛЬТАТИ ІДЕНТИФІКАЦІЇ СПОРІДНЕНOSTІ ВІТЧИЗНЯНИХ ОСВІТНІХ СПЕЦІАЛЬНОСТЕЙ	200
ДОДАТОК Д. АКТИ ВПРОВАДЖЕННЯ РЕЗУЛЬТАТІВ ДОСЛІДЖЕННЯ ..	202

ВСТУП

Обґрунтування вибору теми дослідження. Останнім часом все гостріше постає проблема якісної та своєчасної експертизи наукових творів. Під науковими творами розуміються рукописи статей, доповідей конференцій, розділів книжок та монографій, дисертації, запити на гранти тощо. Експертна оцінка (peer review) вважається ключовою в сучасному управлінні науковими дослідженнями. Стратегічна мета експертизи – це забезпечення якості наукових досліджень шляхом відбору кращих наукових творів, відхилення псевдонаукових робіт, зменшення дезінформації та плутанини, покращення досліджень в результаті виправлення виявлених слабких місць та недоліків.

Останні дослідження вказують на значний запит наукової спільноти щодо покращення експертного оцінювання. Покращення потребує як якість експертних висновків, так і процедура оцінювання. В пріоритеті скорочення тривалості рецензування з огляду на пришвидшення усіх суспільних процесів з відповідними ризиками втрати актуальності тематики або змін у персональному складі наукової команди. Обсяги експертної роботи колосальні; вони часто сконцентровані у дуже стислі часові інтервали. Наприклад, в 2023 р. на конкурс проєктів наукових досліджень і розробок для університетів та наукових установ МОН України подано 749 запитів. Кожен запит оцінюється 5 рецензентами, що сумарно становить біля 3800 експертиз з урахуванням того, що деякі рецензії відхиляються конкурсною комісією через етичні чи змістовні зауваження. За умовами конкурсу усі підготовчі, експертні та узагальнюючі процедури здійснюються протягом 2 місяців. У такому цейтноті за ручного управління експертною діяльністю досягти високої якості висновків надто складно, про що свідчать численні негативні відгуки авторів запитів на отримані рецензії. Намагання забезпечити високу якість експертизи в подібних умовах призводить до значного затягування оцінювання. Наприклад, на конкурс EURIZONE «Remote Research Grants for Ukrainian Researchers» подано понад 750 заявок. Тривалість експертизи заявок вже перевищила планову на 9 місяців і

результатів ще досі немає. Інколи, експертиза затягується і за значно меншої кількості заявок. Наприклад, конкурс «Кембридж – НФДУ 2022», який адмініструвався Національним фондом досліджень України, проведено із запізненням на 4.5 місяці через проблеми з експертизою, хоча подано лише 49 заявок. Аналогічні експертизні проблеми виникають і під час оцінювання рукописів статей, переважно, в редколегіях топових журналів та програмних комітетах престижних наукових конференцій.

Ключовим етапом експертного оцінювання є підбір рецензентів відповідно до тематики твору. Враховуючи постійне зростання кількості наукових творів та потребу скорочення термінів експертизи для економії часу і ресурсів, доцільно максимально автоматизувати підбір рецензентів. При цьому, слід забезпечити: 1) своєчасність підбору рецензентів; 2) незалежність від авторів рецензованої роботи; 3) відповідність рецензента тематиці роботи та багато інших обмежень та вимог. Складнощі автоматичного підбору рецензентів обумовлено необхідністю обробки неструктурованої або слабкоструктурованої інформації, зокрема різноманітних природномовних текстів, які генеруються рецензентами та авторами рецензованих робіт. Результативні наукові дослідженнями в цій проводять Nihar Shah з Carnegie Mellon University, Tang Jie з Tsinghua University, Michal Rosen-Zvi з IBM Research, Steven Jecmen з Carnegie Mellon University, Maryam Karimzadehgan з University of Illinois, але проблематика настільки складна, що вдається вирішити лише окремі частини задачі.

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

Дослідження здійснювалося відповідно до наукового напрямку кафедри комп'ютерних систем управління Вінницького національного технічного університету, зокрема, під час виконання держбюджетної теми Міністерства освіти і науки України 46Д388 «Ідентифікація прихованих залежностей в онлайн-соціальних мережах на основі методів нечіткої логіки та комп'ютерної лінгвістики», номер держреєстрації 0117U000575 та кафедральної

науково-дослідної роботи 46-K7 «Розробка методів і технологій автоматизації процесів контролю і керування». Дослідження виконано з використанням інформаційної системи Dimensions, доступ до відповідних сервісів якої надано Digital Science & Research Solutions Inc. згідно до проєкту DIM-371.

Мета і задачі дослідження.

Мета – підвищення ефективності управління процесами експертизи наукових творів за рахунок розробки інформаційної технології, яка дозволяє автоматизувати процес підбору релевантних рецензентів. Для досягнення цієї мети сформульовано наступні задачі дослідження:

1. Проаналізувати сучасні підходи автоматизованого підбору рецензентів, виявити їх обмеження, переваги та недоліки.
2. Розробити критерій оцінювання схожості потенційного рецензента та наукового твору, які представлені у формі категоріальних розподілів за науковими спеціальностями, частина з яких є спорідненими.
3. Розробити модель розрахунку рівня спорідненості спеціальностей та ідентифікувати за нею поточні рівні спорідненості спеціальностей за доступними статистичними даними.
4. Розробити експрес-метод категоризації рецензента на основі обмеженої інформації про кандидата, зокрема, за коротким списком його наукових інтересів сформованих у вільному стилі.
5. Розробити метод підбору колективу рецензентів з урахуванням якості рішення та обчислювальної складності перебору варіантів.
6. Розробити програмне забезпечення інформаційної технології автоматичного підбору рецензентів та провести експериментальні дослідження для оцінювання практичної цінності запропонованих рішень.

Об'єкт дослідження – інформаційні системи підтримки прийняття рішень для управління експертизою наукової діяльності.

Предмет дослідження – моделі та алгоритми обробки тексту з профілів наукових творів та рецензентів, та прийняття рішень про призначення рецензентів.

Методи дослідження. У дослідженнях використовувались методи теорії реляційних баз даних для формалізації моделей та операцій над даними рецензентів та науковими творами; методи прикладної статистики, комп'ютерної лінгвістики та інтелектуального аналізу даних для розробки методу категоризації ключових слів; методи об'єктно-орієнтованого програмування для розробки програмних складових інформаційної системи, методи дискретної оптимізації для налаштування параметрів алгоритму категоризації та вирішення задачі про призначення колективу рецензентів; методи планування експерименту для формування датасетів та проведення експериментів.

Наукова новизна отриманих результатів.

1. Удосконалено критерій Чекановського для оцінювання схожості категоріальних розподілів шляхом введення додаткового доданку, який враховує перехресну спорідненість категорій, що дозволяє взяти до уваги спорідненість наукових спеціальностей потенційного рецензента та наукового твору.

2. Вперше розроблена модель статистичного оцінювання спорідненості наукових спеціальностей, що ґрунтується на розрахунку індекса Жаккара за базою категоризованих наукових публікацій, що дозволяє швидко ідентифікувати рівень схожості спеціальностей в порівнянні з трудомісткими аналогами на основі аналізу списку цитованості та списку джерел кожної окремої статті.

3. Вперше розроблено метод категоризації рецензентів, який базується на перетворенні обмеженого набору початкової інформації у формі наукових інтересів рецензента або ключових слів заявки у вектор в просторі наукових спеціальностей за принципом word2vec, і відрізняється від формального співставлення шифрів спеціальностей аналізом фактичного змісту заявки та поточної компетентності рецензента, а від моделі співставлення ключових слів – врахуванням синонімії та спорідненості.

4. Вперше розроблено метод призначення команди рецензентів за їх відповідністю тематиці заявки, на відміну від ізольованого підбору рецензентів враховуює здатність саме колективу рецензентів спільно оцінити заявку за усіма аспектами її тематики, та є збалансованим за критеріями якості підбору і витратами ресурсів на пошук колективу.

Практичне значення результатів.

На основі отриманих наукових результатів розроблена інформаційна технологія, яка автоматизує складні та трудомісткі процеси підбору рецензентів з використанням обмеженого набору початкової інформації. Окрім того, ідентифіковано рівні спорідненості наукових спеціальностей в межах Австралійсько-Новозеландської системи класифікації наук ANZSRC-2020 та рівні спорідненості вітчизняних науково-освітніх спеціальностей, що має окремо практичне значення для вирішення таких задач управління освітньо-науковою діяльністю, як корекція переліку спеціальностей та галузей знань, відслідковування трендів міждисциплінарних досліджень, формування перспективних міждисциплінарних освітніх програм та конкурсних пропозицій для широкого конкурсу, модернізації мережі факультетів та університетів. Також сформовано датасети для категоризації науковців та підбору рецензентів, що має окремо практичне значення для перевірки та удосконалення нових алгоритмів обробки такої інформації. Датасети [1], [2] виставлено у публічний доступ. На сформованому датасеті для підбору рецензентів в результаті вдалося покращити склад разових рад дисертацій, при порівнянні з радами від закладу, в середньому на 32-55% в залежності від типу алгоритму, що використовувався.

Особистий внесок здобувача. Усі наукові результати дисертаційної роботи отримано здобувачем самостійно. У працях, опублікованих у співавторстві, здобувачеві належить: в [3], [4], [5], [6], [7], [8], [9], [10], [11] – розробка моделі статистичного оцінювання схожості між двома науковими спеціальностями, запропоновано використовувати індекс Жаккара, проведено аналіз даних за різні періоди; в [12], [13], [14], [15], [16] – розробка алгоритму

категоризації ключових слів в межах деякої системи класифікації наук; в [17] – оптимізація підбору параметрів алгоритму категоризації; в [18], [19] – розробка критерію оцінювання схожості двох категоріальних розподілів враховуючи перехресну спорідненість категорій, в [20] – огляд задач та аналіз методів автоматичного підбору рецензентів.

Апробація матеріалів дисертації. Основні наукові та практичні результати роботи доповідалися та обговорювалися на таких наукових конференціях:

VIII міжнародна науково-технічна конференція «Комп’ютерне моделювання та оптимізації складних систем» (Дніпро, 2023 р.);

II міжнародна науково-практична конференція «Прикладні аспекти сучасних міждисциплінарних досліджень», Донецький національний університет імені Василя Стуса (Вінниця, 2023 р.);

The Sixth International Workshop on Computer Modeling and Intelligent Systems, CMIS 2023 (Запоріжжя, 2023 р.);

XVI міжнародна конференція «Контроль і управління в складних системах» (Вінниця, 2022 р.);

The Fourth International Workshop on Computer Modeling and Intelligent Systems, CMIS-2021 (Запоріжжя, 2021 р.);

Наукова конференція професорсько-викладацького складу, наукових працівників і здобувачів наукового ступеня за підсумками науково-дослідної роботи за період 2019–2020 рр., Донецький національний університет імені Василя Стуса (Вінниця, 2021 р.);

II всеукраїнська науково-практична конференція студентів, аспірантів та молодих вчених «Прикладні інформаційні технології», Донецький національний університет імені Василя Стуса (Вінниця, 2021 р.);

I Науково-технічна конференція підрозділів Вінницького національного технічного університету, (Вінниця, 2021 р.);

XLIX Науково-технічна конференція підрозділів Вінницького національного технічного університету, (Вінниця, 2020 р.);

7th International Conference «Information Technology and Interactions» (Київ, 2020 р.);

XV міжнародна конференція «Контроль і управління в складних системах» (Вінниця, 2020 р.);

DCSMart'2019: 1st International Workshop on Digital Content & Smart Multimedia (Львів, 2019).

Публікації. Основні результати дисертаційного дослідження викладені у 18 наукових працях. Опубліковано 4 статті у фахових журналах України зі спеціальності 126 [9], [10], [18], [20]; 5 статей, що включені до наукометричної бази даних Scopus [3], [4], [12], [13], [19]; 9 тез доповідей на міжнародних та науково-практичних конференціях [5], [6], [7], [8], [11], [14], [15], [16], [17].

Документи інтелектуальної власності, що отримані за результатами дисертації.

Отримано авторське свідоцтво на комп'ютерну програму [21].

Структура та обсяг дисертації. Дисертація складається зі вступу, чотирьох розділів, висновків, списку використаної літератури зі 172 найменувань та 5 додатків. Загальний обсяг дисертації становить 207 сторінок, з них 139 сторінок основного тексту, який містить 62 рисунки та 10 таблиць.

РОЗДІЛ 1

АНАЛІЗ СТАНУ ПРОБЛЕМИ ТА ПОСТАНОВКА ЗАДАЧ ДОСЛІДЖЕННЯ

1.1 Аналіз процесів експертизи наукових творів як об'єкту дослідження

Останнім часом все гостріше постає проблема якісної та своєчасної експертизи наукових творів. Під науковими творами розуміються рукописи публікацій (статей, тез доповідей, розділів монографій, монографій), дисертації, запити на грантові дослідження тощо. Експертне оцінювання (в іноземній літературі *peer review*) вважається основним інструментом оцінювання наукових досліджень в сучасному світі [22], [29]. Експертне оцінювання у своєму поточному вигляді набуло широкого загалу лише в середині двадцятого століття [30]. Експертне оцінювання аналізує наукові твори з точки зору компетентності, значущості та оригінальності [31]. Його основна мета – забезпечення якості наукових робіт для зменшення дезінформації та плутанини [32] та покращання якості робіт перед їх опублікуванням [33].

Опитування науковців з різних галузей знань показують, що експертне оцінювання високо цінується багатьма науковцями [22]. Ці опитування також вказують на те, що є значний і зростаючий запит щодо покращення процесу експертного оцінювання. Покращення потребують як якість експертних висновків, так і процедура оцінювання. В пріоритеті скорочення тривалості рецензування з огляду на пришвидшення усіх суспільних процесів з відповідними ризиками втрати актуальності тематики або змін у персональному складі наукової команди. Обсяги експертної роботи колосальні, при чому вони часто сконцентровані у дуже стислі часові інтервали. Наприклад, в 2023 р. на конкурс проєктів наукових досліджень і розробок для університетів та наукових установ МОН України подано 749 запитів. Кожен запит оцінюється 5 рецензентами, що сумарно становить біля 3800 експертиз з урахуванням того,

що деякі рецензії відхиляються конкурсною комісією через етичні чи змістовні зауваження. За умовами конкурсу усі підготовчі, експертні та узагальнюючі процедури здійснюються протягом 2 місяців. У такому цейтноті за ручного управління експертною діяльністю досягти високої якості висновків надто складно, про що свідчать численні негативні відгуки авторів запитів на отримані рецензії. Намагання забезпечити високу якість експертизи в подібних умовах призводить до значного затягування оцінювання. Наприклад, на конкурс EURIZONE «Remote Research Grants for Ukrainian Researchers» подано понад 750 заявок. Тривалість експертизи заявок вже перевищила планову на 9 місяців і результатів ще досі немає. Інколи, експертиза затягується і за значно меншої кількості заявок. Наприклад, конкурс «Кембридж – НФДУ 2022», який адмініструвався Національним фондом досліджень України, проведено із запізненням на 4.5 місяці через проблеми з експертизою, хоча подано лише 49 заявок.

За даними двох престижних конференцій кількість поданих заявок за 6 років збільшилась приблизно в 5 разів (рис. 1.1). Більш масштабні дослідження показують, що щорічний приріст опублікованих статей становить 4.1% з подвоєнням кожні 17.3 роки [34]. Аналіз здійснювався за 4 бібліометричними базами даних: Dimensions, Microsoft Academic, Web of Science та Scopus. Це показує важливість експертного оцінювання у сучасному світі та необхідність швидкого пошуку експертів які можуть надати таку оцінку.

Експертне оцінювання також має і недоліки та ключові проблеми, які потребують вирішення. Припускається, що експертне оцінювання надає механізм для раціонального, чесного та об'єктивного прийняття рішень [22]. Для цього необхідно забезпечити незалежність оцінювання від соціальних особистостей авторів та рецензентів, та від теоретичних упереджень рецензента [35]. Однак, для досягнення цих цілей виникає кілька проблем. Експертне оцінювання є системою в якій основна роль виконується людиною. Відповідно можливі прояви людського фактору. Тому експертне оцінювання буває упереджене та неефективне; інколи можуть діяти і корупційні мотиви або

плагіат ідей [36]. Окрім цього є також проблема з популярністю авторів, оцінювання робіт яких відбувається під впливом ефекту Матвія (Matthew effect). У [37] було проведено дослідження, в результаті якого виявилось що експертне оцінювання сильно упереджене щодо авторів, у яких є багато опублікованих робіт, що зменшує шанси інших не менш талановитих авторів на позитивне рішення. Це показує що задача експертного оцінювання є досить складною і може часто не виконувати свою роль належним чином.

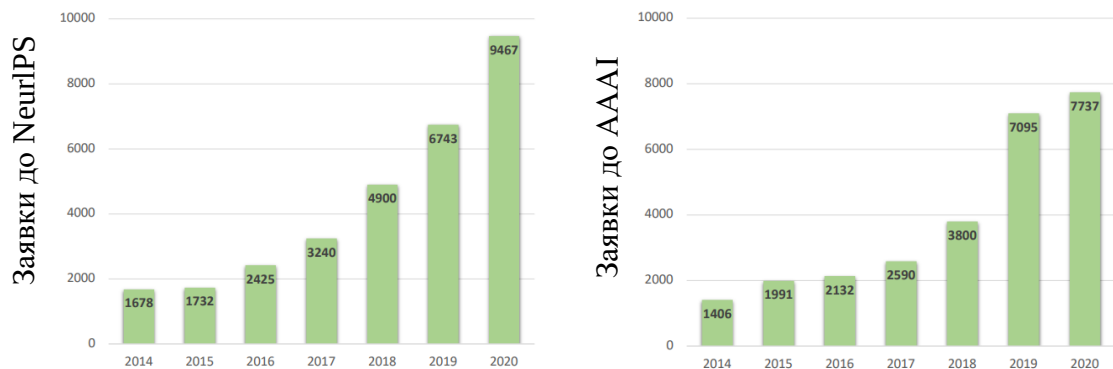


Рисунок 1.1 – Розподіл кількості поданих заявок на дві престижні міжнародні наукові конференції за останні кілька років [22]

На рис. 1.2 подано спрощену схему процесу експертного оцінювання рукопису журнальної публікації. Вона складається з декількох етапів. Після подачі роботи редактор перевіряє її на предмет відповідності до встановлених вимог; якщо знайдено невідповідність то заявка відхиляється. Далі виконується підбір рецензентів та рецензування. Заявку можуть прийняти без змін, прийняти з виправленням недоліків або відхилити.

Ключовим етапом експертного оцінювання є підбір рецензентів відповідно до тематики твору. Цю задачу виконують вручну, автоматично або ж у змішаному режимі. Враховуючи постійне зростання кількості наукових творів доцільним є виконувати це автоматично для економії часу і ресурсів. Тому в цій роботі ми аналізуємо існуючі підходи до автоматичної рекомендації рецензентів для експертизи наукових творів.

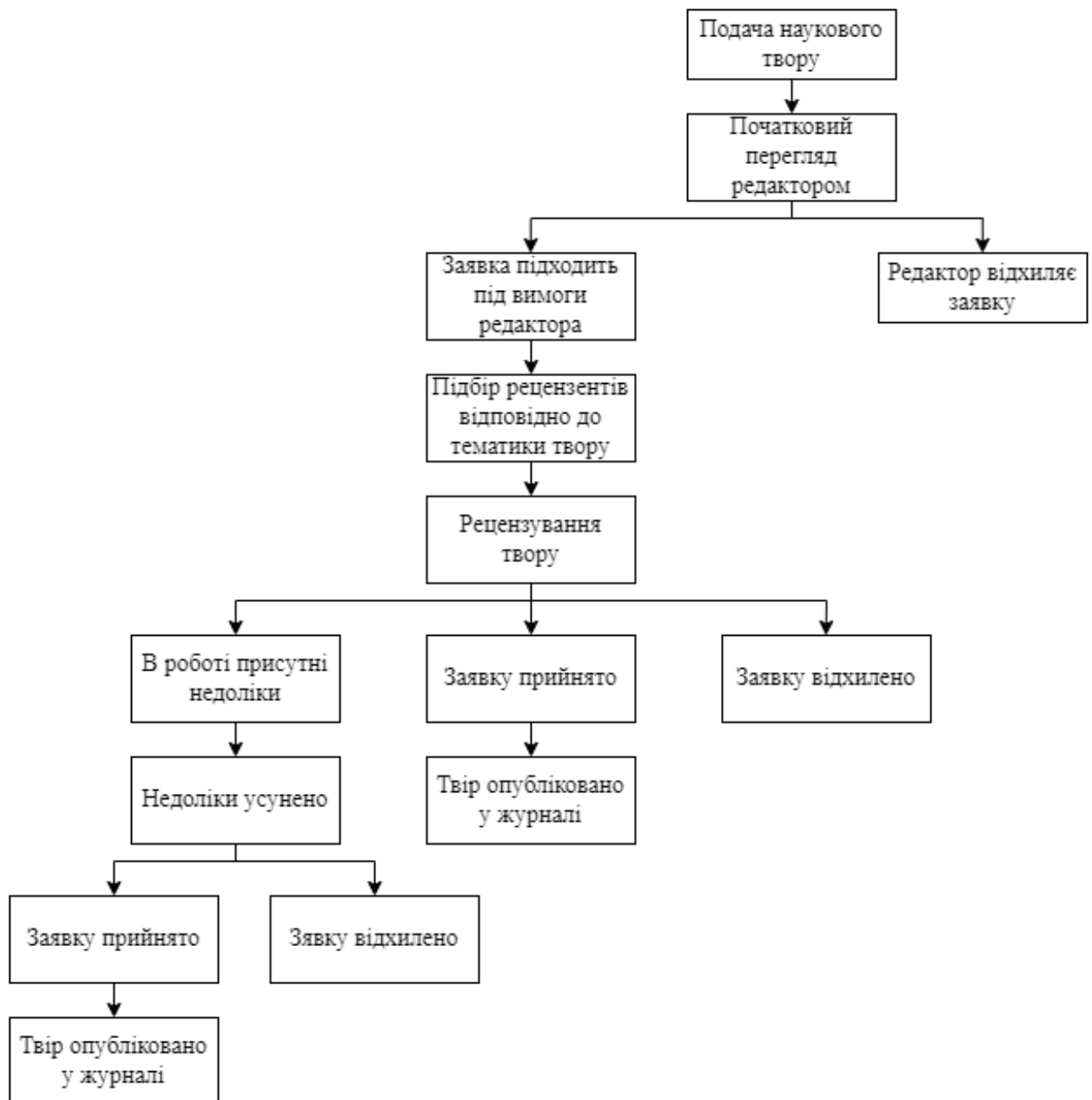


Рисунок 1.2 – Спрощена схема експертного оцінювання рукопису журнальної статті [38]

1.2 Задача підбору рецензента для надання експертної оцінки

Задача підбору рецензента для надання експертної оцінки (в іноземній літературі *Reviewer assignment problem*) є важливою задачею для сучасної академічної спільноти. Окрім згаданого в попередньому підрозділі стрімкого росту кількості поданих статей до публікації, на це також вказує і кількість

опублікованих досліджень на цю тему та зацікавленість дослідників у ній. Приміром у [39] та [40] нещодавно були проведені літературні огляди на цю тематику, та проаналізовано відповідно 103 та 67 наукових публікацій. Кількість публікацій у Scopus зі словосполученням «*Reviewer Assignment*» у назві або в ключових словах в 2019-2023 рр. становить 66, проти 29 в 2014-2018 рр.. Найперші рішення для автоматичного підбору рецензентів представлено в 1992 р. [39].

Задача призначення рецензентів може розглядатись як частинний випадок узагальненої задачі про призначення (*generalized assignment problem*) [39]. Пошук підходящих висококласних рецензентів є критичним для виявлення наукових досліджень потенційно корисних з практичної точки зору а також для покращення ефективності академічної діяльності [41], [42]. В літературі можна знайти різноманітні роботи по підборі рецензентів для експертної оцінки різних видів робіт: підбору партнерів для спільних досліджень [28]; дослідницьких та розробницьких проєктів [43], [44], [45]; підбору експертів для установ [46], [47]; підбору дослідницьких проєктів для студентів та підбору наукових керівників аспірантам [48]. Рекомендація і підбір рецензентів повинні бути адекватними та ефективними так, як це напряму впливає на репутацію науковців, конференцій та журналів [40].

Система експертного оцінювання розглядається як основний спосіб забезпечення якості наукових публікацій [49] і має потенціал покращити дослідження [50]. Під час експертної оцінки подані статті аналізуються групою експертів – рецензентами, які оцінюють завершеність, важливість і оригінальність дослідження, і рекомендують статтю до публікації або ж до відмови у публікації. Одна з найбільших проблем в експертній оцінці це виявлення та підбір підходящих рецензентів чий дослідницький профіль співпадає з темою поданої заявки [46]. Незважаючи на прогрес в технологіях ці задачі часто виконуються вручну редакторами або організаторами конференцій.

Ручний підхід до підбору рецензентів для експертного оцінювання статей має очевидні недоліки. Зазвичай процес починається з наукових комітетів

конференцій чи журналів, які отримують бази даних науковців для пошуку рецензентів з найбільш підходящим профілем для кожної статті. Це безумовно є часозатратною операцією, особливо коли кількість рецензентів у базах даних та кількість поданих статей є великою. Окрім цього, за ручного підходу зазвичай рекомендують науковців у такій же науковій галузі, як і рецензована стаття. Проте багато інших факторів і індикаторів окрім наукової галузі повинні розглядатись при підборі рецензентів, зокрема, аналізуватися публікації, наукові проєкти та патенти. Ручний підхід ігнорує можливі зв'язки авторів статті з рецензентами, що може призвести до конфлікту інтересів. Враховуючи необхідність оперувати великою кількістю інформації про рецензентів та статті на основі людського суб'єктивного критерію є також ризик виникнення упередження під час підбору рецензентів [40].

Враховуючи всі вище перелічені проблеми є необхідність застосування інтелектуальних технологій для аналізу та отримання важливої інформації з документів та неструктурованих текстів для автоматичного підбору найбільш підходящих рецензентів для оцінки наукових творів. Різноманітні підходи були запропоновані для вирішення задачі підбору рецензентів, включаючи рекомендаційні системи, системи підтримки прийняття рішень [42], [51], [52], підходи на основі машинного навчання [53], лінійного програмування [54], [55], [56] та багато інших. Susan Dumais та Jacob Nielsen [57] вперше розробили автоматичну систему на основі штучного інтелекту використовуючи інформацію про рецензентів з HYPERTEXT'91 конференції у 1991 р. Вони використовували латентний семантичний індекс (Latent Semantic Index) для порівняння анотацій рецензентів з анотаціями поданих статей, сортуючи статті від найменш до найбільш схожих з рецензентами. Але результати були гіршими за ті, які отримано за ручним підходом. Не зважаючи на це, стаття вважається першим дослідженням, яке надало поштовх до розробки нових підходів задачі автоматизації підбору рецензентів [40].

Без прив'язки до того, як структурований процес підбору рецензентів, автори з [58] виділяють три етапи цього процесу.

1. Пошук рецензентів а також вибір методу представлення даних про рецензентів та заявок.

2. Підрахунок схожості між заявкою та рецензентами.

3. Оптимізація співвідношення заявок та рецензентів, щоб максимізувати показник схожості за певних обмеженнях (баланс навантаження на рецензентів, кількість заявок, які необхідно прорецензувати, вподобання рецензентів).

В літературі [59], [60], [61], [62], [63] етапи 2 та 3 поділяють на два можливих типа: а) задача підбору рецензента на основі інформаційного пошуку (information retrieval-based reviewer assignment problem); б) задача підбору рецензента на основі оптимізації (optimization-based reviewer assignment problem). Класифікація цих задач подана на рис. 1.3.

Інформаційний пошук акцентує увагу на другому етапі задачі підбору рецензента, де основна мета – визначення коефіцієнту схожості між рецензентом та заявкою [57], [64]. В цьому випадку задача підбору рецензента є задачею інформаційного пошуку. Відповідно до методів інформаційного пошуку кожна заявка розглядається як запит, а рецензент як документ. В залежності від типу інформації і способу її використання цю категорію методів можна поділити на такі групи: методи на основі тексту/змісту (текст – лінгвістичні особливості, зміст – інформація яка подається в тексті); методи на основі мереж (графи); методи на основі ручного підбору. Ручні методи підбирають рецензентів на основі додаткової інформації, яку надають самі ж рецензенти, наприклад, теми які вони хотіли б рецензувати. Це зазвичай вимагає ручного збору інформації.

Методи на основі тексту/змісту зазвичай використовують методи обробки природньої мови, зокрема статистику співпояви інформації у документах, наприклад між словами та темами (тематична модель), між словами та документами (мовна модель), та між словами (моделі ембедингу слів). Методи на основі мереж зазвичай використовують мережеві зв'язки, такі як мережа цитувань, мережа співавторства, мережа співпраць та інші пов'язані мережі між рецензентом і рецензентом та між рецензентом і заявкою.

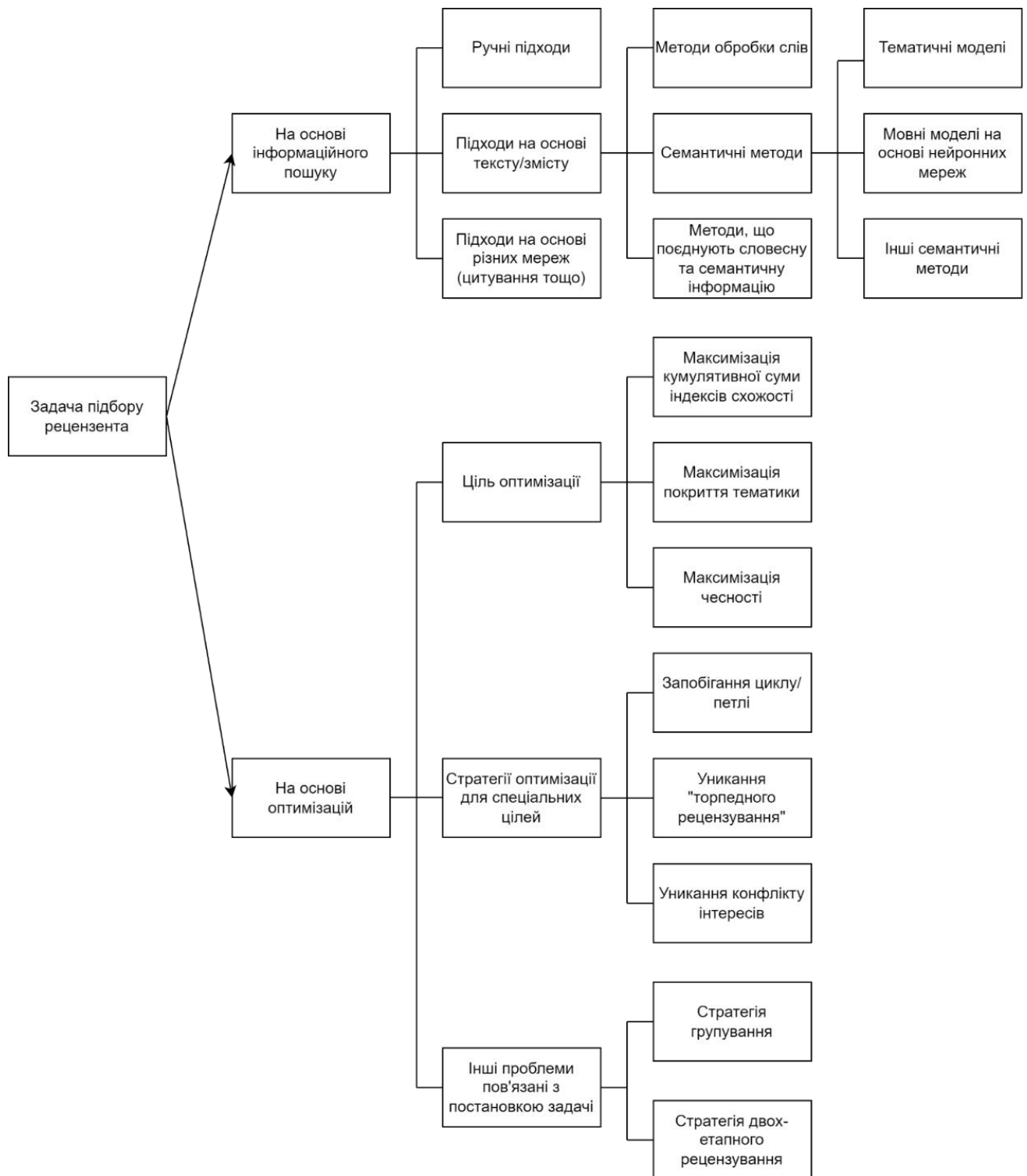


Рисунок 1.3 – Класифікація підходів до розв’язання задачі підбору рецензента [39]

Задача підбору рецензента на основі оптимізації зосереджена на третьому етапі підбору рецензента, і перетворює інформаційний пошук на оптимізаційну задачу, яка вирішується з точки зору дослідження операцій [60], [65], [66], [67].

1.3 Аналіз даних, що використовуються для побудови профілів рецензента та заявки

Автоматичний підбір рецензентів передбачає наявність інформації про рецензентів, на основі якої можна співставити їх до відповідних заявок. Це відноситься до першого етапу підбору рецензента. Науковці використовують декілька видів інформації одночасно для представлення рецензента та заявки. Представлення даних про рецензента та заявку називають профілем. Досить часто для побудови профіля використовується така інформація про опубліковані статті рецензента, як назва, анотація, ключові слова, повний текст, список посилань, список цитувань [39]. Додатково до цих даних, профілі можуть включати також особисту інформацію про рецензентів: ім'я та прізвище, місце роботи, область інтересів, область експертизи, *h-index*, мережа цитувань, інформація про співавторства, інші зв'язки (наприклад, академічні стосунки, колеги, конкуренти, співробітники на проєктах, співпраці тощо), репутація та статус, минулий досвід рецензування, досвід подачі заявок, досвід управління проєктом (кількість проєктів, якість проєктів) тощо [39].

На рис. 1.4 подано розподіл кількості статей по видам інформації, що використовується для формування профіля заявки на рецензування. Як видно, лідерами є анотація, повний текст, ключові слова та назва. Для найточнішого підбору рецензента необхідно використовувати якомога більше інформації про нього та про заявку, але це в свою чергу ускладнює сам процес прийняття рішення. Тому необхідно віднайти баланс між складністю та зрозумілістю.

На рис. 1.5 подано розподіл кількості статей по видам інформації зі статей, що використовується для побудови профіля рецензента. Як видно, лідерами є повний текст, анотація, назва та ключові слова. Сюди також входить рік публікації, кількість публікацій рецензента та посилання на інші роботи.



Рисунок 1.4 – Розподіл кількості статей по видам інформації, що використовується для побудови профілю заявки [39]

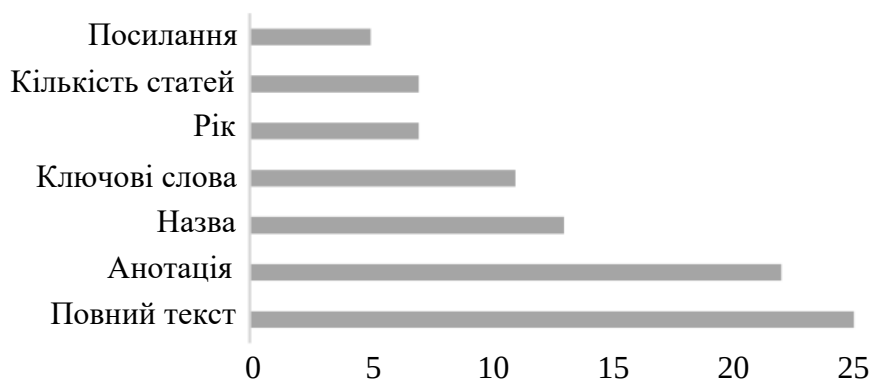


Рисунок 1.5 – Розподіл кількості статей по видам інформації статей рецензентів, що використовується для побудови профілю рецензента [39]

На рис. 1.6 подано розподіл кількості статей по типам інформації що використовується для побудови профіля рецензента. Найбільш використовувані види інформації це галузь дослідження / галузь експертизи, інформація про співавторство та місце роботи. Основною причиною вибору саме цієї інформації є її легкодоступність [39]. Цю інформацію зазвичай надають самі рецензенти, але її може бути замало, щоб в повній мірі описати профіль рецензента. Окрім цього, кандидати можуть помилятися при вказанні своєї експертизи або галузі дослідження. З цієї причини в літературі також використовують інформацію про опубліковані роботи рецензентів.



Рисунок 1.6 – Розподіл кількості статей по видам інформації, що використовується для побудови профілю рецензента [39]

1.4 Аналіз методів обробки даних про рецензентів та заявки

У підр. 1.2 розглянуто етапи розв'язання задачі підбору рецензента. Нижче розглядаються деякі аспекти моделювання природньої мови для побудови представлення даних про рецензентів та заявок на рецензування.

Як відомо, комп'ютер оперує лише числами і єдиний спосіб для нього зрозуміти природну мову – перетворити текст у числа. Методи для виконання даної операції розроблялись ще з початку виникнення напрямку обробки природньої мови (Natural Language Processing (NLP)) [27], [68], [69], [70], [71], [72], [73], [74], [75]. До найпопулярніших відносять:

- 1) сумка слів (bag-of-words);
- 2) частота терму - зворотна частота документу (TF-IDF – term frequency-inverse document frequency) [75];
- 3) позитивна матрична факторизація [76];
- 4) латентне розподілення Діріхле [77];

- 5) латентний семантичний аналіз [78];
- 6) тематичне моделювання;
- 7) розподілене представлення слів і фраз та їхня композиційність (word2vec) [50];
- 8) глобальні вектори для представлення слова (glove) [72];
- 9) двонаправлений енкодер представлення на основі перетворювачів (BERT) [68];
- 10) глибокі контекстні представлення слова (ELMo) [69].

1.4.1 Сумка слів

Даний метод представлення тексту у формі зрозумілій для комп'ютера є найпростішим і полягає у представленні відповідних елементів тексту їх індексами із словника. У випадку представлення слова це буде індекс слова у словнику, що створений із слів тексту, що опрацьовується. Цілий словник використовується для створення багатовимірному простору у якому кожний документ перетворюється у вектор частоти терму (слова) через підрахунок появи термів у документі.

Векторизація тексту є основою для багатьох сучасних алгоритмів обробки тексту [79]. Однак, використання термів як елементарних одиниць має свої недоліки. Для підкреслення обмеження представлення сумки слів використаємо приклад з [79] порівнюючи схожість між наступними трьома реченнями:

1. Obama speaks to the media in Illinois;
2. The president greets the press in Chicago;
3. A dinosaur is in exhibition at Chicago meseum.

Перші два з цих речень мають досить схоже значення навіть враховуючи те, що різні слова були використані для опису ймовірно однієї і тієї ж події. Однак, звичайна евклідова відстань між частотними векторами термів не може цього визначити, вона не враховує схожості між значеннями цих слів. Приведенні до нижнього регістру, перетворенні і лематизованні речення подано у табл. 1.1, а косинус подібності між трьома векторами подано у табл. 1.2.

Відповідно до табл. 1.2 речення 2 і 3 є ближчими ніж 1 і 2. Це показує важливість значення слів під час автоматичного аналізу та використання для різних NLP-задач.

Таблиця 1.1 – Частота термів фраз з прикладу

Слова	Речення		
	1	2	3
at	0	0	1
chicago	0	1	1
dinosaur	0	0	1
exhibition	0	0	1
greet	0	1	0
illinois	1	0	0
in	0	1	1
media	1	0	0
meseum	0	0	1
obama	1	0	0
president	0	1	0
press	0	0	0
speak	0	0	0
to	1	0	0

Таблиця 1.2 – Матриця косинусів подібності серед векторизованих фраз

	Vec_1	Vec_2	Vec_3
Vec_1	0	1	1
Vec_2	1	0	0.592
Vec_3	1	0.592	0

1.4.2 TF-IDF

На противагу сумці слів, що враховує лише частоту появи певного елементу текста у документі, у [80] запропоновано підхід, що буде представлення елементів тексту на основі добутку двох значень: частоти появи елемента в певному документі колекції та логарифму зворотної частоти документів колекції (кількість документів поділена на кількість документів в яких зустрічається елемент тексту). Такий підхід дозволяє відсіяти стоп-слова, що мають велику частоту появи, і при цьому підсилити рідко вживані слова. Це є значним покращенням простого підходу на основі частоти, але це все рівно не дає можливості виконувати семантичне порівняння слів.

1.4.3 Word2Vec

Суть цього методу [73] полягає у використанні так званої skip-gram моделі – ефективного методу для навчання високоякісних векторних представлень слів з великої кількості не структурованої текстової інформації [81]. Даний метод базується на нейронних мережах і на відміну від своїх аналогів більш ефективний у часі, тому що не використовує повного перемноження матриць. Ціль - навчити векторне представлення слова на основі його появи в контексті інших слів (рис. 1.8).

Особливістю цієї моделі є те, що просте додавання векторів може утворити інше представлення слова. Наприклад, $\text{vec}(\text{"Ukraine"}) + \text{vec}(\text{"river"})$ близьке до $\text{vec}(\text{"Dnipro River"})$, а $\text{vec}(\text{"Germany"}) + \text{vec}(\text{"capital"})$ близьке до $\text{vec}(\text{"Berlin"})$. Це свідчить про те що, певний рівень розуміння мови може бути отриманий використовуючи базові математичні операції над векторами представлення слів.

Мета навчання skip-gram моделі – знайти представлення слова, що добре відображають навколишні слова у реченні або документі. У більш формальному визначенні за наявності слів $w_1, w_2, \dots, w_T$, мета моделі це максимізація середньої логарифму ймовірності появи слова в контексті навколишніх слів:

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log(p(w_{t+j}, w_t; \theta)), \quad (1.1)$$

де c – розмір навчального контексту (що може бути функцією від слова посередині w_t). Чим більше c тим більше навчальних прикладів і відповідно більша точність за рахунок більшого часу навчання.

Рис. 1.9 показує як word2vec-модель в деякій мірі вирішує проблему значення слів, що описана вище. З рисунку видно, що споріднені слова розташовані ближче у просторі (тут використано метод зменшення просторовості для зручності представлення даних).

1.4.4 Glove

Glove це ще один метод представлення тексту, дослівно розшифровується як глобальні вектори для представлення слів. На відміну від word2vec, цей метод базується на факторизації матриць, що формуються з контексту слова [72]. Спочатку формується інформація про спів-появи, тобто для кожного слова рахується частота його появи у деякому контексті у великій базі тексту. Словам відповідають рядки матриці, а контексту – стовпчики. Метод заснований на спостереженнях із табл. 1.3 де зв'язок між словами *ice* і *steam* досліджується шляхом знаходження відношення їх співпояв у деякому контексті k . Для слів k споріднених з *ice* але не з *steam* (наприклад, $k = \text{solid}$) очікується, що відношення їх співпояв з k буде велике. І навпаки, для слів k пов'язаних з *steam* але не з *ice* (наприклад, $k = \text{gas}$), співвідношення очікується малим.

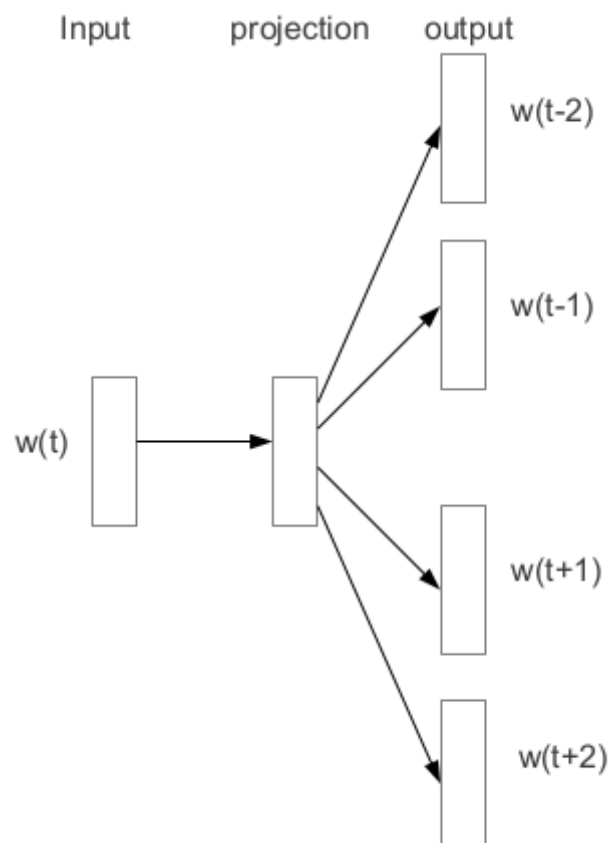


Рисунок 1.8 – Архітектура skip-gram моделі [81]

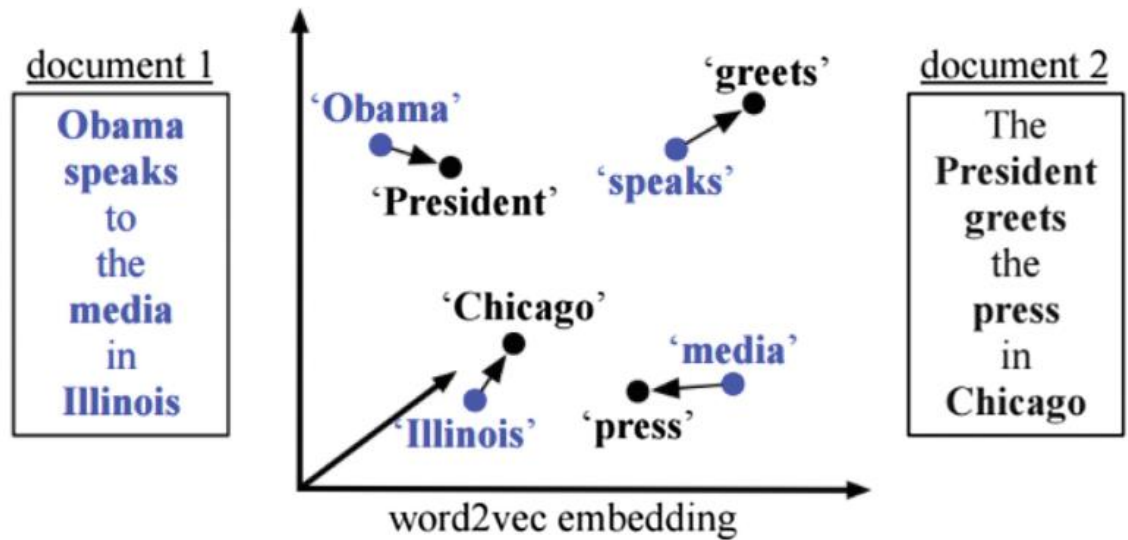


Рисунок 1.9 – Семантично схожі слова у просторі word2vec [79]

Таблиця 1.3 – Ймовірності співпояви з обраним контекстом з бази тексту розміром 6 мільярдів токенів [72]

Ймовірність і співвідношення	$k=solid$	$k=gas$	$k=water$	$k=fashion$
$P(k ice)$	$1.9 \cdot 10^{-4}$	$6.6 \cdot 10^{-5}$	$3.3 \cdot 10^{-3}$	$1.7 \cdot 10^{-5}$
$P(k steam)$	$2.2 \cdot 10^{-5}$	$7.8 \cdot 10^{-4}$	$2.2 \cdot 10^{-3}$	$1.8 \cdot 10^{-5}$
$P(k ice)/P(k steam)$	8.9	8.5×10^{-2}	1.36	0.96

Утворені вектори слів за допомогою glove допомагають встановити значення слова у різноманітних контекстах. Деякі важливі властивості цих векторів:

1. Чудово справляються з задачею пошуку аналогії. Зв'язок між словами визначається відстанню між ними;

2. Так само як і word2vec дозволяє утворювати нові вектори представлення додаючи існуючі, що є близькими до ймовірних слів представлення.

1.4.5 Тематичне моделювання

Тематичне моделювання це метод, що представляє документ у просторі меншої розмірності а ніж сумка слів оцінюючи вектор у просторі тем (topics) [77]. Тематичне моделювання часто використовується як інструмент для видобування інформації з тексту (text-mining tool), для знаходження прихованих латентних структур у корпусі тексту, однак, розмірність і компоненти кожного вектору, представленні як теми, не обов'язково корелюють із змістом, який людина може очікувати. До алгоритмів тематичного моделювання відносять: позитивна матрична факторизація [82]; латентний розподіл Діріхле (LDA) [77]; латентний семантичний аналіз (pLSA) [78]. Найпопулярніший з них - латентний розподіл Діріхле.

У тематичному моделюванні кожний документ представляється як ймовірнісний розподіл по усіх темах, а кожна тема це ймовірнісний розподіл по усьому словнику слів з тексту.

1.5 Аналіз методів підбору рецензентів на основі інформаційного пошуку

У підр. 1.2 розглянуто етапи розв'язання задачі підбору рецензента. У даному підрозділі розглядається поточний стан розв'язання другого етапу задачі, а саме підрахунку індексу схожості між заявками та рецензентами. Таку процедуру в літературі називають задачею підбору рецензента на основі інформаційного пошуку. Виділяють 3 підходи до розв'язання другого етапу [39]: а) ручний; б) на основі тексту/змісту; в) на основі нейронних мереж.

1.5.1 Ручні підходи

Ручні підходи використовують інформацію, яку надають рецензенти для встановлення відповідності між заявками та рецензентами. Найбільш поширеним методом є рейтингування, в якому рецензент заздалегідь оцінює заявку для рецензування на основі особистих вподобань [39], [44], [46], [51], [56], [60], [67], [83], [84], [85], [86], [87], [88], [89], [90], [91], [92]. Цей метод

передбачає, що рецензенти проглядають анотації або повний текст заявки і призначають бали для кожної заявки. Бали інтерпретуються як готовність рецензента оцінити конкретну заявку. Кінцевим результатом рейтингування є список для рецензентів з їхніми вподобаннями по відношенню до заявок. Цей список і представляє собою індекс схожості між рецензентом та заявками. Чим вищий бал вподобання, тим вищий індекс схожості. За такої схеми рецензент сам вирішує чи є у нього достатній рівень експертизи, щоб оцінити певну заявку. За правильного застосування рейтингування добре сприймається рецензентами, воно покращує якість і надійність оцінювання за умови, що бали виставлено чесно і відповідально [93]. Рейтингування має й очевидні недоліки. Головний з них це те, що за великої кількості заявок рейтингування сильно ускладнюється і займає багато часу. Інший недолік полягає у тому, що багато конференцій мають нестачу необхідних балів для багатьох заявок [39]. Якщо під час попереднього оцінювання заявок більшість з них не отримали балів, то результуюча матриця рейтингування буде розрідженою, що значно знизить точність підбору рецензента. Окрім цього, рецензенти схильні давати оцінку своєї відповідності заявці з точки зору своїх поточних дослідницьких інтересів, а не з точки зору всього їхнього досвіду. Рейтингуванням також вразливе до недоброчесної поведінки рецензентів, що може призвести до конфлікту інтересів, а також вплинути на результат експертизи. Приміром, у випадку коли автори заявок виступають і рецензентами, вони можуть підвищити свої шанси бути призначеними до певної заявки шляхом стратегічного виставлення балів на заявки яким вони хочуть дати погану оцінку і таким чином підвищити рейтинг своєї заявки [91], [94].

Інші ручні способи підбору рецензента використовують ключові слова з якоїсь певної експертної тематики, що надаються рецензентами та авторами заявок [95], [51], [55], [88], [90], [96], [97], [98]. Під час подачі заявок від авторів вимагають надати короткий список ключових слів, що описує їхню роботу. В той самий час рецензенти подають список ключових слів, що характеризує їхню експертизу. Ключові слова можуть обиратись з якогось певного словника

або бути не обмеженими у виборі словника. Отже, заявка і рецензент представляються у вигляді списку ключових слів. Найпростіший спосіб поррахувати схожість між цими списками – це поррахувати кількість ключових слів що є у обох списках одночасно [95], [88]. Також використовують індекс Жаккара на основі спільних ключових слів [90], [98].

Ручні підходи можуть бути ефективними за малої вибірці даних, а також коли експертизу рецензента та заявки можна описати ключовими словами з точним семантичним значенням. За необхідності підбору великої кількості рецензентів цей метод втрачає свої переваги. Окрім цього, метод незахищений віж недоброчесної поведінки рецензентів.

1.5.2 Підходи на основі аналізу тексту/змісту

Підходи які обробляють текст (публікації, ключові слова тощо) в автоматичному режимі використовують методи обробки природньої мови. Під обробкою тексту в даному випадку мають на увазі обробка лінвістичних особливостей, під змістом мається на увазі обробка самої суті інформації, яка подається в тексті, тобто її значення. Підходи на основі тексту/змісту є підходами на основі інформаційного пошуку, за якими підбирають рецензентів порівнюючи характеристики тексту з профілю рецензента з характеристиками тексту з профілю заявки [99]. Ці підходи є автоматичними методами, які знаходять неявні схожості без додаткових дій і обчислюють індекс схожості між рецензентами та заявками використовуючи технології інтелектуального аналізу тексту. В літературі їх зазвичай поділяють на 2 типи: методи на основі слів та семантичні методи [29], [100].

Методи на основі слів використовують статистичні методи для підрахунку частоти появи слів. Вони розділяються на дві категорії: моделі векторного простору та статистичні мовні моделі.

Моделі векторного простору представляють кожену заявку і рецензентів як вектори сумки слів помножених на TF/IDF індекс даних з профілей заявок та рецензентів. Вони представляють кожне слово як числовий вектор, який відповідає тому як слово використовується. Маючи таке представлення даних

про заявки та рецензентів, задача визначення схожості між ними зводиться до порівняння двох векторів. Перша така модель була запропонована у 1999 р. в [101]. В ній індекс схожості між рецензентом та заявкою розраховано за метрикою схожості векторів, а саме за косинусом подібності. У [84] автори порівнюють 3 підходи на основі моделей векторного простору для підрахунку індексу схожості: TF/IDF представлення повного тексту документів; добування ключових слів з документів на основі машинного навчання для представлення заявки та рецензента як вектора ключових слів; онтологічне добування тематик для точного представлення заявок та рецензентів як тематичних векторів (тематичні вектори це ті які визначені над деякою множиною тем, тема являє собою об'єднану сукупність деяких понять та ідей які часто зустрічаються у тексті). Результати на основі трьох підходів показали, що використання більшої кількості інформації надає кращі співставлення профілів заявок та рецензентів в більшості випадків. TF/IDF також використовувався і в інших дослідженнях. В деяких з них одночасно використовувалися і тематики, які отримані з тексту на основі різних методів [102], [103] та [104]. Є також дослідження де TF/IDF не використовувався безпосередньо для підрахунку схожості, але лише для отримання попередньої інформації про текст, що потім використовувалась в інших методах обробки природньої мови, наприклад з використанням нечіткої логіки [105].

Моделі векторного простору мають декілька недоліків, і основний з них це те, що вони базуються на лексичному співпадинні і припускають, що слів достатньо для того, щоб однозначно тлумачити зміст тексту. Проте це спричиняє втрату контексту а також помилки при визначенні схожості у випадку слів близьких за значенням, але різних за написанням.

Статистичні мовні моделі за своєю суттю подібні до моделей векторного простору тим, що вони також намагаються описати текст з профілів заявок та рецензентів використовуючи математичні моделі, але роблять це у інший спосіб. Вони використовують статистичні та ймовірнісні підходи для пошуку

ймовірнісного розподілу частин тексту таких як слова, речення тощо. Роботи з цими моделям проведено у [106], [107], [107], [109].

Головна проблема методів на основі слів полягає в тому, що вони використовують частоту появи слів в тексті профілів заявки та рецензента, і не враховують семантичне значення тексту. Це стосується синонімічних або близьких за значенням виразів, де, наприклад, вирази «нечітка множина» та «нечіткий висновок» можуть бути взагалі несхожими з точки зору цих моделей. Також це має місце і для полісемічних виразів, тобто тих які мають декілька значень в залежності від контексту. Тому підрахунок схожості між заявкою та рецензентом в деяких випадках може давати хибні результати. Для вирішення цієї проблеми в літературі пропонуються семантичні моделі визначення схожості між заявками та рецензентами.

Семантичні моделі визначення схожості, зазвичай, засновані на тематичному моделюванні. Тематичне моделювання це метод, що представляє документ у просторі меншої розмірності, аніж сумка слів, а саме у просторі тем. У літературі зазначено використання різних методів тематичного моделювання для задачі підбору рецензентів, зокрема прихованого семантичного індексування [57], ймовірнісного прихованого семантичного аналізу [108], [109], прихованого розподілу Діріхле [77], моделі «Автор – Тема» [61], [97], [110], [111], [112], моделі «Автор – Персона – Тема» [64], моделі «Автор – Дисципліна – Тема» [113], моделі «Автор – Інтерес–Тема» [114].

Приховане семантичне індексування (Latent semantic indexing) запропоновано у 1990 р. [115]. Цей метод дозволяє знайти тематичний розподіл слів використовуючи метод сингулярного розкладання матриці. Перше дослідження з використанням цього методу проведено у [57] – автори представляли тематичний розподіл профілів заявок та рецензентів як точки у k -вимірному просторі і за схожість обрали косинус подібності цих точок. У [116] створено систему підбору рецензента використовуючи приховане семантичне індексування для підрахунку схожості між заявками та рецензентами. Ця система покращила якість підбору рецензентів на одній з конференцій.

Приховане семантичне індексування має декілька недоліків: для точних результатів необхідно мати велику кількість даних; низька інтерпретабельність через зменшення розмірності; вимагає багато обчислювальних потужностей; вектори займають багато місця. Як покращення цих недоліків запропонований ймовірнісний прихований семантичний аналіз (Probabilistic latent semantic analysis) [78], який надає кращу інтерпретацію прихованого виміру векторів, ніж семантичне індексування, а також дає кращі результати у випадку полісемії та синонімії [117].

Як покращення попередніх тематичних моделей у [77] запропонована модель прихованого розподілу Діріхле, яка вважається однією з найкращих сучасних моделей тематичного моделювання. В ній документи представляються випадковою сумішю прихованих тематик, що знаходиться на основі ймовірнісного моделювання спільної появи (co-occurrence) слів у тексті. Вона застосовувалася у таких дослідженнях з підбору рецензентів як [55], [63], [64], [97], [100], [107], [111], [118], [119], [120], [121], [122], [123], [124], [125].

Розглянемо варіації прихованого розподілу Діріхле для задачі підбору рецензента. В [24] запропоновано статистичну модель «Автор – Тема» на основі тематичного моделювання з використанням прихованого розподілу Діріхле. Модель представляє дані про потенційного рецензента як розподіл над деякими абстрактними темами. Теми є кластерами схожих слів. Її недоліком є погана інтерпретація тем, оскільки вони формуються за частотою слів у одному документі. Для покращення інтерпретабельності у [64] для підбору рецензентів запропонована модель «Автор – Персона – Тема». В ній враховано те, що автори часто пишуть про декілька різних комбінацій тем з однієї предметної області. Дуже рідко особа є експертом в усіх аспектах якоїсь предметної області. За результатами моделювання науковця розділяють на декілька так званих персон (personas). Кожна персона є кластером статей науковця із своїм тематичним розподілом. У [114] запропоновано модель «Автор – Інтерес – Тема», яка є узагальненням моделей «Автор – Тема» та «Автор – Персона –

Тема», та на відміну від них дозволяє моделювати профіль рецензента використовуючи декілька прихованих випадкових змінних одночасно.

У [113] запропонована модель «Автор – Дисципліна – Тема» (рис. 1.10). У ній для опису науковця додатково використовують наукову спеціальність, яка визначається за журналом чи збірником статей, у якому опублікована аналізована робота. Таким чином результуюча модель буде мати інформацію, яка характеризує спільні області дослідження різних документів. Їхня головна ідея полягає у тому, що експертиза науковця може бути представлена у ієрархічній формі, де коренем є деяка макро-концепція, а листками є конкретні види контенту, що пов'язані з концепцією у корені. Для перевірки цієї моделі автори використали базу даних у галузі Information Systems and Management. За результатами дослідження цієї бази даних виявлено, що дана модель дає кращі результати, аніж модель «Автор – Тема».

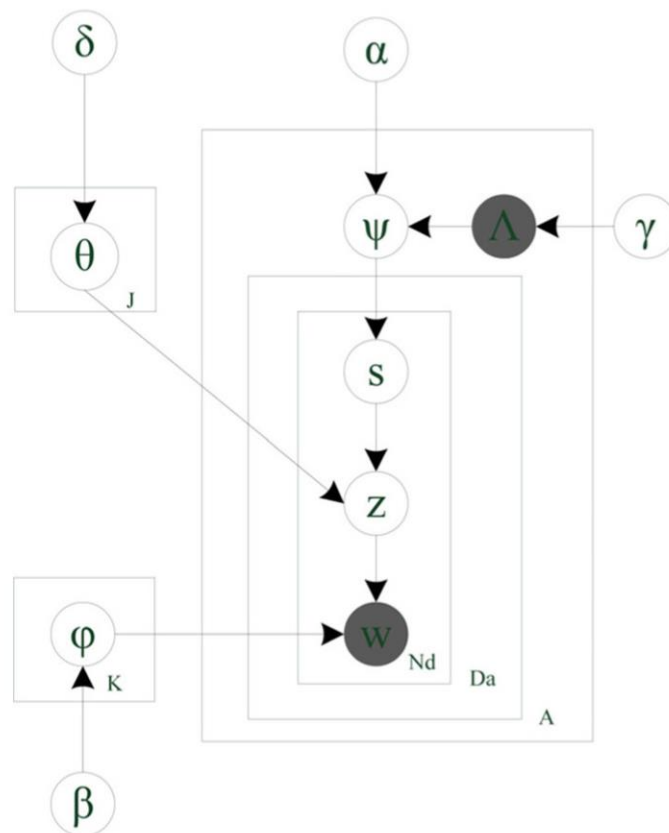


Рисунок 1.10 – Модель «Автор – Дисципліна –Тема» [113]

У [126] запропоновано ще одне покращення моделі «Автор – Тема» шляхом включення у модель даних по самій конференції. Модель має назву «Автор – Конференція». Хоча значення певної тематики залишається незмінним, кореляція різних тематик може змінюватися з часом. На основі цієї ідеї у [127] було запропоновано модель, що дозволяє визначати рецензентів певної тематики для різних періодів і показувати як інтереси та зв'язки рецензентів змінювались з часом.

Хоча теми в прихованому просторі вважаються семантично зв'язними, тематичні моделі будують розподіл спільної появи слів в тексті, а не семантичної інформації цих слів [128]. Окрім цього тематичні моделі вимагають великої кількості текстових даних для знаходження ефективного розподілу, а це може бути проблемою, коли, наприклад, для моделювання доступна лише анотація [129]. Також припускається, що слова з аналізованого тексту є незалежними та некорельованими. Це значно покращує ефективність обробки текстів, але ігнорує семантичні зв'язки між словами.

1.5.3. Підходи на основі нейронних мереж

Розглянуті раніше методи для представлення тексту з профілів заявки та рецензента за своєю суттю є частотними. Вони розглядають документи як множину слів ставлячи акцент на частоті кожного слова, але при цьому не враховується порядок та взаємозв'язок сусідніх слів. Отримати семантичні представлення тексту з цими методами неможливо. Частотні методи на великих об'ємах тексту породжують розріджені вектори, що відноситься до так званого «прокляття розмірності», коли зі зростання розмірності вектору він стає все розрідженішим. Для вирішення цих проблем в літературі пропонується використання мовних моделей на основі глибинного навчання та нейронних мереж. Такі моделі утворюють так звані вектори ембедингу (word embedding), які представляють семантичні особливості елементів тексту. Слово ембедінг в даному випадку означає вбудовування слова у числовий неперервний простір шляхом перетворення його у вектор чисел таким чином, що просторові зв'язки цих векторів у даному просторі відображають семантичні зв'язки між словами.

Використовуючи ембединг текст можна представляти у відносно маловимірному просторі. Зазвичай моделі ембедингу поділяють на два типи: статичні та контекстуальні. Статичні моделі можуть кодувати морфологічні і семантичні характеристики тексту. До цих моделей відносять Word2Vec [73], Glove [72], FastText [130]. Контекстуальні моделі ембедингу включають в себе глибокі нейронні мережі ULMFiT [131], BERT [132], GPT-2 [133], DistillBert [134]. На основі ембедингу пропонуються методи для вирішення задачі підбору рецензента. Розглянемо деякі з них.

У [79] поєднують тематичне моделювання, word2vec та так звану метрику mover distance для визначення подібності науковців на основі їх анотацій у статтях для вирішення задачі пошуку партнерів для спільних досліджень, що є дуже схожою до задачі підбору рецензента. За результатами дослідження система є успішною у визначенні існуючих спів-авторств з вибірки даних не зважаючи на те, що ця інформація не була використана під час навчання. Система пропонує валідні потенційні можливості для співпраці із споріднених галузей дослідження незалежно від попередньої співпраці. На рис. 1.11 подано схематичне представлення цієї моделі. Подібність науковців визначається на основі подібності їх публікацій, подібність публікацій визначається на основі подібності тематичного розподілу публікації, подібність тематичного розподілу визначається на основі подібності слів.

У [135] запропонована модель CCRес, яка поєднує інформацію про публікації науковців та мережу співпраць між ними для створення кращої рекомендації науковців для спільних досліджень, що можна узагальнити і на задачу підбору рецензента. Загалом автори зробили наступний внесок: 1) створили модель CCRес, що поєднує методи представлення науковця на основі його даних і на основі його взаємодії з іншими науковцями; 2) запропонували так звані характеристичні вектори (feature vectors), що показують академічні особливості науковців у різних галузях знань, ця модель використовує модель тематичної кластеризації та модель випадкової ходьби (random walk) для обрахунку впливу науковця у кожній галузі науки; 3) проведено експерименти

на даних з DBLP-датасету для оцінки продуктивності CCRес за різноманітних сценаріїв. Сама модель базується на припущенні, що науковець зазвичай хоче співпрацювати з іншими науковцями, які мають такі ж наукові інтереси і сильний вплив у науковій спільноті. Модель тематичної кластеризації та розділення науковців є методом на основі контенту, що генерує різні області і відображає науковців у ці області.

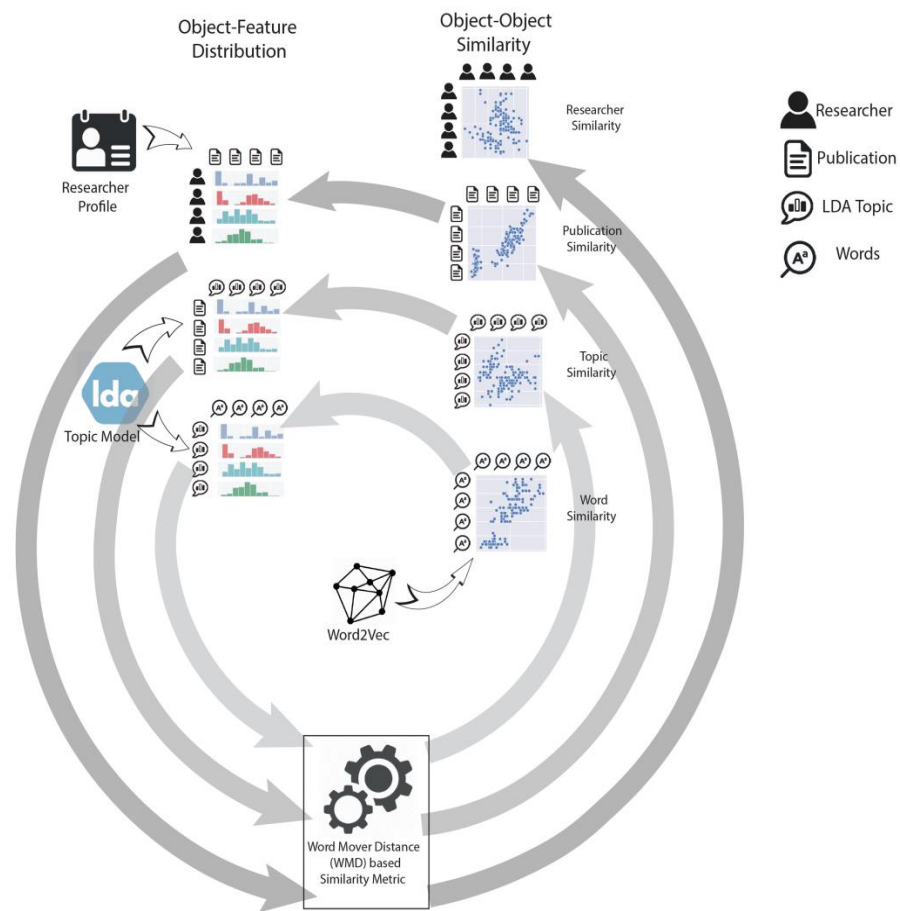


Рисунок 1.11 – Схема моделі для знаходження подібності між науковцями [79]

Автори моделі CCRес вирізняють декілька методів обробки тексту включаючи вище згадані word2vec [73], прихований розподіл Діріхле (LDA) [77] та прихований семантичний аналіз (pLSA) [78]. Однак автори зупиняються на моделі word2vec тому, що вона підходить найбільше у даному випадку застосування. LDA та pLSA породжують ймовірнісний розподіл слів та документів базуючись на спільній появі слів та документів, що фокусуються на

описі конотативних (з побіжними значеннями) тематик. word2vec у свою чергу на контекстуальній інформації (семантичній та синтаксичній) слів. Для моделі CCRес вхідними даними для кластеризації є назви усіх статей науковця. Хоча вони і стислі, але є повними реченнями і містять контекстуальну інформацію, яку word2vec може використати для породження більш точного опису характеристик науковця. За результатами дослідження модель CCRес показала, що поєднання методів на основі контенту та взаємодії між науковцями може генерувати значно кращі рекомендації науковців для співпраці. Однак тут є ще місце для досліджень. Використання заголовків статей містить значно менше інформації ніж анотація чи ключові слова.

У [136] запропонована обмежена згорткова нейронна мережа (A Restricted Convolutional Neural Network for Expert Finding EFCCN) для пошуку експертів. Автори використали модель згорткової нейронної мережі, що добре себе зарекомендувала у задачах розпізнавання зображень. Задачу пошуку рецензента можна звести до порівняння документів: документу, що рецензується та документу ймовірного рецензента. Схожість між двома документами подається у вигляді квадратної матриці, що складається з схожості пар слів з різних документів. Однак така матриця ігнорує специфічність слів та залежить від позиції слів. Для цього для кожної пари слів автори також використали метрику оберненої частоти документу (IDF). Таким чином, задачу визначення схожості між рецензентом та рецензованою роботою зведено до задачі розпізнавання зображень. Слова з документів у цій моделі також представляються за допомогою word2vec [73]. Однак використання згорткових мереж для такої задачі має свої недоліки, передусім через те що, матриця схожості є розрідженою та просторово переривчастою, що погано впливає на продуктивність мереж. З огляду на це автори модифікували оригінальний згортковий шар нейронної мережі, що має сприяти успішному навчанні. За результатами досліджень автори показують, що методи на основі нейронних мереж, що використовують ембедінг слів можуть краще навчатись для представлення даних про рецензентів. Однак як і будь які нейронні мережі, ця

модель має погану інтерпретабельність і високу складність. Висока складність також впливає на швидкість роботи алгоритму. Оскільки нейронні мережі навчаються, це вимагає досить великої кількості даних, що робить їх використання практично неможливим через відсутність відкритих даних про рецензентів різних робіт та вимагає додатковий збір даних, що ймовірноше повинна виконувати людина.

У [129] запропонований PaRe-підхід для підбору рецензентів до статей за простором спільних тематик. Цей підхід моделює тематики, що характерні для статті та профілю рецензента на основі абстрактних тематичних векторів. Одним з недоліків підходу до представлення науковців є те, використовуються ключові слова статей. Ключові слова автори можуть вказувати довільним чином, хтось може вказувати одне ключове слово, хтось десятки. Це в свою чергу призводить до труднощів під час моделювання представлень. В [129] вирізняють 5 видів моделей для моделювання представлення рецензентів та заявок:

1) моделі на основі ознак, які представляють документи для рецензування у вигляді сукупності ключових слів, що характеризують тематику роботи;

2) ймовірнісні моделі, наприклад, прихований розподіл Діріхле [77], у якому кожна тема представлена як ймовірнісний розподіл по деякому словникові слів, а документ є сукупністю прихованих тематик;

3) моделі ембедингу, в яких документ представляється як один вектор;

4) графові моделі, які використовують інформація про співавторство у формі графа, у якому вершини відповідають рецензентам, ребро вказує про наявність роботи рецензентів над спільною публікацією, вага ребра визначається кількістю публікацій;

5) моделі на основі нейронних мереж, в яких семантичне представлення документу є результатом навчання нейронної мережі шляхом порівняння векторів документів рецензента та документа для рецензії.

Система PaRe включає декілька етапів підбору рецензентів. На етапі збирання даних знаходяться статті з конференцій у тій ж самій галузі, що й

робота для рецензування. Ці дані використовуються для створення пулу кандидатів в рецензенти. Джерело даних – Microsoft Academic Graph. Усі анотації рецензента об'єднуються в один документ, який потім використовується для створення профілю рецензента, який відображає його наукові тематики. Етап обробки даних базується на використанні ембедингу слів, які перед-навчаються (pre-train) моделлю CBOW з word2vec [73]. Етап пошуку рецензента використовує звичайний метод моделювання тематик. Модель порівнює анотації роботи, що рецензується, з анотаціями потенційних рецензентів. За результатами тестування модель PaRe перевищує за точністю всі інші існуючі моделі рекомендації рецензентів. Одним із недоліків такої моделі є погана інтерпретація породженого профілю науковця і також не врахування зміни наукових інтересів для визначення доречності належності рецензента до певної тематики на основі його нещодавніх вкладень у певну область науки.

Оскільки як заявка, так і рецензент, можуть відноситись до деяких галузей досліджень, то в літературі також розглядають задачу підбору рецензента як задачу багатовірликової класифікації [137]. У [137] запропоновано модель Ніераг, яка категоризує заявку та рецензента до деяких дослідницьких галузей, а далі схожість визначається шляхом порівняння дослідницьких галузей.

1.6 Аналіз методів для підбору рецензентів на основі оптимізації

На третьому етапі задачі підбору рецензента виконується безпосереднє співставлення рецензентів з відповідними заявками. На другому етапі відбувалась обробка даних профілів рецензента та заявки для того, щоб їх можна було семантично порівнювати. Результатом другого етапу є представлення профілів рецензента та заявки у вигляді числових векторів, які можна порівнювати і знаходити схожість між ними використовуючи метрики схожості між векторами. Третій етап використовує ці вектори для того, щоб

визначити сумісність пари «рецензент-заявка» з точки зору різних критеріїв. Тобто третій етап зводиться до умовної оптимізації за деякої цільової функції. В літературі інколи це також називають узагальненою задачею призначення (generalized assignment problem) [39], у якій кожному об'єкту деякої множини потрібно поставити у відповідність один або декілька об'єктів іншої множини. Маючи множину рецензентів та заявок, а також рівень схожості між кожною парою «рецензента-заявка», задачу можна представити як оптимізацію на повному дводольному графі [39], [138].

Виділяють такі цільові функції для оптимізації: максимізація суми індексів схожості рецензентів та заявок [46], [66], [67], [96], [107], [139]; максимізація покриття тематики [44], [54], [110], [111], [118], [138], [140], [141]; оптимізація справедливості [87], [95], [142], [143], [144].

Як критерій оптимальності можна обрати суму усіх індексів схожості між профілями призначених рецензентів та заявок. Для цього необхідно мати матрицю схожості між профілями рецензентів та заявок, що знаходиться на другому етапі задачі підбору рецензента. Це вважається найпростішим способом оптимізації призначення рецензентів до заявок.

Щоб виконати таку оптимізацію в літературі пропонують використати декілька різних підходів. Найчастіше зводять оптимізацію до задачі цілочисельного лінійного програмування (integer linear programming), яка використовує лінійну цільову функцію та обмеження [96], [139], [148], [149], [150], [151], [152]. Це спрацьовує коли кількість рецензентів та заявок є невеликою, оскільки задача цілочисельного лінійного програмування є NP-повною. Тому в літературі пропонуються варіанти її вирішення на основі апроксимації та евристичних алгоритмів. До таких відносять різні варіації жадібних алгоритмів [45], [55], [66], [109], [110], [144], [146], генетичні алгоритми [60], [153], [154], метод рою частинок [63] та інші. Також використовують модель дводольного графа, за якою задачу підбору рецензента зводять до задачі пошуку максимального потоку на графі [46] або до задачі пошуку потоку мінімальної вартості [155].

Як критерій оптимальності обирають і покриття тематики. Вважається, що на другому етапі вирішення задачі підбору рецензента профілі рецензента та заявки представлені як деякі тематичні вектори, тобто кожний елемент вектору відповідає певній тематиці. Тематикою, наприклад, може бути якась галузь знань чи наукова спеціальність. Цілком ймовірно, що рецензент так само як і заявка можуть відноситись одночасно до декількох тематик, і тому є сенс щоб співпадіння між тематиками рецензента та заявки було максимальним. Для цього дослідники пропонують різні підходи до оптимізації покриття тематики [44], [110], [111], [118], [138], [140], [141]. Поміж них є також методи, які додатково враховують завантаженість на рецензентів та мінімальну допустиму кількість оцінок заявки [54], [110].

Два попередні підходи оптимізації призначення рецензентів до заявок в цілому дають адекватні результати, однак може виникати також проблема справедливості. Під справедливістю розуміється те, що експертний рівень рецензентів рівномірно розподілена між заявками. Під час оптимізації суми індексів схожості чи покриття тематики може виникати ситуація коли рецензенти з високою відповідністю призначаються до певної підмножини заявок, а інші заявки оцінюють рецензентами з низьким рівнем експертності. Це несправедливо по відношенню як рецензентів – вони працюватимуть за різних умов комфортності. Це несправедливо і для авторів заявок – за одними заявками будуть отримані достовірні оцінки, а за іншим – достовірність оцінок буде низькою через слабе володіння рецензентами тематикою заявки. Найпростіший спосіб забезпечити чесність – це додати обмеження до цільової функції оптимізації про те, що заявка повинна мати принаймні одного рецензента з певним рівнем експертизи [95]. Більш сучасні підходи пропонують на додаток до загальної цільової функції та умов оптимізації додати умову «локальної справедливості», тобто кожна заявка повинна бути призначена до групи рецензентів що мають достатньо експертизи для надання оцінки [143].

Окрім згаданих вище обмежень, застосовують і інші. До найпопулярніших обмежень відносять [39]:

- завантаженість на рецензента – рецензент може оцінити лише певну кількість робіт;
- мінімальна кількість експертних оцінок заявки;
- уникнення конфлікту інтересів – рецензент не призначається до заявки, якщо виявлено які-небудь зв'язки між ним та авторами заявки – спільне місце роботи, співавторство інших робіт, відносини «ментор-студент», робота над спільним проектом тощо;
- захист від стратегічної конкурентності (strategyproofing) – у випадку коли рецензентами є самі автори, рецензенти можуть штучно занижувати оцінки іншим з метою покращення статусу своїх робіт [25], [145];
- уникання «торпедного рецензування» — рецензенти можуть прагнути бути призначеними до конкретних заявок, які їм не подобаються, з метою надання їм поганих оцінок [91];
- запобігання циклу/петлі – у випадку коли рецензентами є самі автори може виникнути ситуація коли 2 рецензенти оцінюють заявки один одного, що може спричинити конфлікт інтересів якщо рецензенти знайомі і намагаються підвищити рейтинг один одного [94], [146], [147].
- експертність рецензентів повинна покривати усі тематики заявки, тобто у рецензентів повинно бути достатньо знань для того щоб оцінити заявку [54], [98], [143].

1.7 Дослідницька ідея та задачі подальшого дослідження

Розглянуті підходи до вирішення задачі автоматичного призначення рецензентів в більшості випадків потребують доволі великої кількості інформації про статті рецензентів, їхню взаємодію з іншими науковцями та аналогічну інформацію про авторів заявок. Такі підходи мають декілька недоліків. По-перше, для збору такої інформації необхідно витратити певний час та інші ресурси, що не завжди є доступним. По-друге, постає питання актуальності експертизи рецензентів у вигляді статей, оскільки статті можуть

застарівати і не відображати поточний напрямок діяльності рецензентів, особливо коли рукопис статті занадто довго не публікувався. Окрім цього, статті можуть писатись кількома авторами і встановити за що саме відповідав конкретний автор у статті є складною задачею. По-третє, виникає питання коли і як робити оновлення даних про рецензентів та авторів заявок. Це все вимагає додаткових людських ресурсів. З іншого боку рецензенти можуть вказати свої наукові інтереси у вигляді короткого набору ключових слів, що характеризуватимуть саме їх область експертизи. Наприклад, як це роблять в Google Scholar. Така початкова інформація буде недостаєм для врахування таких критеріїв як конфлікт інтересів, але її можна використати для швидкого експрес-підбору певної кількості рецензентів, що підходять до тематики дисертації і таким чином значно спростивши пошуки релевантних кандидатів. В подальшому, цих рецензентів можна детальніше аналізувати на предмет виявлення конфліктів інтересів чи досягнення інших критеріїв оптимізації. Такий підхід досягає балансу за критеріями якості підбору і витрати ресурсів на формування колективу.

Щодо методів аналізу текстів, то пропоновані варіанти в літературі є або занадто примітивними, або занадто складними. У першому випадку можуть виникати проблеми з визначенням семантичної схожості, а у другому випадку постає проблема складності навчання моделей, а також розуміння їх результатів для кінцевого користувача. Тому виникає проблема балансування між коректністю та інтерпретованістю результатів моделювання тексту. Для цього нами пропонується виконувати категоризацію ключових слів заявки та рецензентів шляхом їх перетворення у вектор в просторі наукових спеціальностей, що значно спрощує інтерпретованість результатів.

На основі вказаних недоліків та обмежень, мета дослідження має наступний вигляд.

Мета дослідження – підвищення ефективності управління процесами експертизи наукових творів за рахунок розробки інформаційної технології

експрес-підбору релевантних рецензентів. Для досягнення цієї мети необхідно вирішити наступні задачі:

1. Розробити критерій схожості потенційного рецензента та наукового твору, які представлені у формі категоріальних розподілів за науковими спеціальностями, частина з яких є спорідненими.
2. Розробити модель розрахунку рівня спорідненості спеціальностей та ідентифікувати за нею поточні рівні спорідненості спеціальностей.
3. Розробити експрес-метод категоризації рецензента на основі обмеженої інформації про кандидата, зокрема, за коротким списком його наукових інтересів сформованих у вільному стилі.
4. Розробити метод підбору колективу рецензентів з урахуванням якості рішення та обчислювальної складності перебору варіантів.
5. Розробити програмне забезпечення інформаційної технології автоматичного підбору рецензентів та провести експериментальні дослідження для оцінювання практичної цінності запропонованих рішень.

Висновки до розділу 1

1. Виконано огляд актуальності та стану розв'язання проблеми автоматичного підбору рецензентів. Встановлено, що задачу підбору рецензентів вирішують у декілька етапів: пошук кандидатів в рецензентів з вибором методу представлення даних про рецензентів та заявок; підрахунок схожості між заявкою та рецензентами; оптимізація призначень заявок та рецензентів, щоб максимізувати деякий критерій відповідності при певних обмеженнях (баланс навантаження на рецензентів, кількість заявок які необхідно прорецензувати, вподобання рецензентів тощо).

2. Розглянуто та проаналізовано різноманітні підходи до обробки природньої мови в контексті задачі підбору рецензентів. Встановлено, що поширені підходи потребують доволі великого обсягу інформації про статті

рецензентів, їх взаємодію з іншими науковцями та аналогічну інформацію про авторів заявок. Це вимагає витра значних ресурсів.

3. Встановлено, що поширені методи аналізу текстів є або занадто примітивними, або занадто складними. Необхідно віднайти баланс між коректністю та інтерпретованістю результатів моделювання тексту. Для цього, нами пропонується виконувати категоризацію ключових слів рецензентів шляхом їх перетворення у вектор в просторі наукових спеціальностей, що значно спрощує інтерпретацію результатів.

4. На основі проведеного огляду стану розв'язання проблеми автоматичного підбору рецензентів сформовано задачі подальшого дослідження.

Результати розділу опубліковано в таких роботах автора: [16], [20].

РОЗДІЛ 2

УЗАГАЛЬНЕНИЙ ОПИС АЛГОРИТМУ ПІДБОРУ РЕЦЕНЗЕНТА

2.1 Загальна постановка задачі експрес-підбору рецензента

Задача підбору рецензента складається з трьох етапів: 1) пошук рецензентів а також вибір методу представлення даних про рецензентів та заявки; 2) підрахунок схожості між заявкою та рецензентами; 3) оптимізація співвідношення заявок та рецензентів. В даній роботі ми припускаємо, що наявний список потенційних рецензентів. Нижче ми фокусуємо увагу на способах обробки даних про заявки та рецензенти, а також на способах призначення рецензентів до заявок. В роботі більша увага приділяється питанням експрес-підбору рецензентів, коли потрібно сильно звужити великий список кандидатів. Подальший короткий список може аналізуватися або вручну, або передаватися на процедуру тонкого підбору, яка є ресурсно-витратною і вимагає наявності значно більшого об'єму початкової інформації, ніж для експрес-підбору.

Під експрес-підбором рецензентів розуміється такий підбір, що враховує лише семантичну схожість між заявками та рецензентами, іншими словами відбувається підрахунок індексів схожості і далі призначається необхідна кількість рецензентів таким чином, щоб забезпечити максимальну відповідність колективу рецензентів до заявки згідно деякого критерію.

Під профілем мається на увазі векторне представлення даних про заявку чи рецензента. Як згадувалось у першому розділі, для побудови профілю заявки та рецензента може використовуватись різна інформація. На першому етапі підбору необхідно обрати дані, що використовуватимуться для прийняття рішення а також метод представлення цих даних у векторній формі. У випадку заявки ми використовуємо список ключових слів, а у випадку рецензента – список ключових слів сформований з наявних про нього даних. Цей список ключових слів може бути як зі свіжих публікацій кандидата, так із його CV чи

профілю з деякого реєстру науковців. У другому випадку, зазвичай, ці 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. Щодо методу представлення даних, то існуючі підходи зазвичай використовують різні статистичні моделі, тематичні моделі, або ж моделі ембедингу. Деякі з них аналізують частоту появи слів у тексті, інші формують вектори представлення на основі спів-появи слів у тексті. Загалом результуючі вектори представлення погано інтерпретуються, а також вимагають великої кількості даних для отримання таких представлень. В цій роботі пропонуємо підхід, який будує представлення ключових слів у векторному просторі шляхом зведення цієї задачі до задачі багаторішкової класифікації ключових слів в межах деякої системи класифікації наук. Таким чином, кінцеве представлення профілів заявки та рецензента буде виглядати як розподіл над деякими спеціальностями науки. Нижче подано повний опис цього процесу класифікації а також сам алгоритм підбору.

2.2 Представлення даних заявки та рецензента

Для представлення даних про заявку та рецензента ми використовуємо багаторішкову класифікацію, яку ще також називають категоризацією (рис. 2.1), тобто віднесення об'єкта з ознаками (атрибутами) $X = (x_1, x_2, \dots, x_n)$ до кількох класів із $\{l_1, l_2, \dots, l_m\}$. З математичної точки зору класифікація – це відображення виду $X = (x_1, x_2, \dots, x_n) \rightarrow y \in \{l_1, l_2, \dots, l_m\}$. У нашому випадку атрибутами заявки є ключові слова, а атрибутами потенційного рецензента є ключові слова сформовані на основі наявних даних про нього. А класами є спеціальності науки визначені в межах деякої системи класифікації наук. Наприклад, наукова стаття може бути віднесена до кількох наукових напрямків (рис.2.2): галузі знань 46 *Information and Computing Sciences* з підгалуззями знань 4605 *Data Management and Data Science* та 4611 *Machine Learning*, і галузі знань 47 *Language, Communication and Culture* з підгалуззю 4704 *Linguistics*.

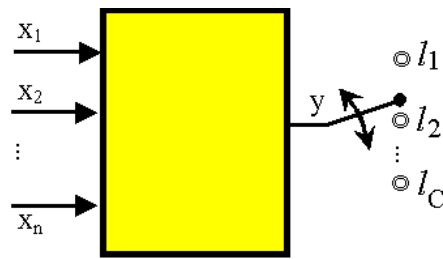


Рисунок 2.1 – Задача класифікації [157]

Publication - Preprint

Language Models are Few-Shot Learners

arXiv, arXiv:2005.14165 - May 2020
<https://doi.org/10.48550/arxiv.2005.14165> / <https://arxiv.org/abs/arXiv:2005.14165>

Authors
 Tom B. Brown
 Benjamin Mann
 Nick Ryder
[28 more](#)

Abstract
 Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions - something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches. Specifically, we train GPT-3, an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model, and test its performance in the few-shot setting. For all tasks, GPT-3 is

[Dynamic building defect categorization through enhanced unsupervised text classification with domain-specific corpus embedding methods](#)
 Kahyun Jeon, Ghang Lee, Seongmin Yang, Yonghan Kim, Seungah Suh
 2024, Automation in Construction - Article
[Get PDF](#) [Add to Library](#)

[Fine-tuning GPT-3 for machine learning electronic and functional properties of organic molecules](#)

[View PDF](#)

[Add to Library](#)

[Share](#)

[Export citation](#)

Publication metrics [About](#)

Dimensions Badge

	218	Total citations
	217	Recent citations
	111	Field Citation Ratio
	n/a	Relative Citation Ratio

Research categories

Fields of Research (ANZSRC 2020)

- 46 Information and Computing Sciences
- 4605 Data Management and Data Science
- 4611 Machine Learning
- 47 Language, Communication and Culture
- 4704 Linguistics

Рисунок 2.2 – Категоризація статті у Dimensions

2.2.1 Високорівневий опис алгоритму категоризації ключових слів в межах системи класифікації наук

Для виконання категоризації необхідно зібрати велику кількість даних зі словами та відповідними словосполученнями, які категоризовані до галузей науки визначених в певній системі, а також навчити модель машинного навчання розпізнавати класи ключових слів. Це потребує великої кількості людських ресурсів для розмічення слів класами, а також постійного оновлення даних тому що можуть з'являтися нові ключові слова специфічні в межах якоїсь галузі науки. Натомість ми пропонуємо використовувати існуючі бази публікацій, які категоризовані в межах деякої системи класифікації наук. Такі бази назвемо базами розмічених документів, в яких кожній публікації співставлено одну або декілька галузей науки чи спеціальностей (рис. 2.2). Таким чином, визначити клас ключового слова можна дізнавшись в яких

публікаціях воно використовувалось, і вже далі отримати список галузей науки до яких ці публікації категоризовано. Схематично це зображено на рис. 2.3. На ньому показано, що в базі розмічених документів кожна стаття категоризована до деяких галузей науки. Наприклад, *Стаття 1* категоризована відповідно до *Науки 1* та *Науки 2*. Беручи до уваги це, можна побудувати розподіл появи ключового слова в контексті різних наук. Наприклад, для ключового слова з рис. 2.3 розподіл появи матиме наступний вигляд: *Наука 1* – 1 поява, *Наука 2* – 1 поява, *Наука 3* – 2 появи. На основі такого розподілу далі виконується категоризація ключового слова в межах системи класифікації наук.

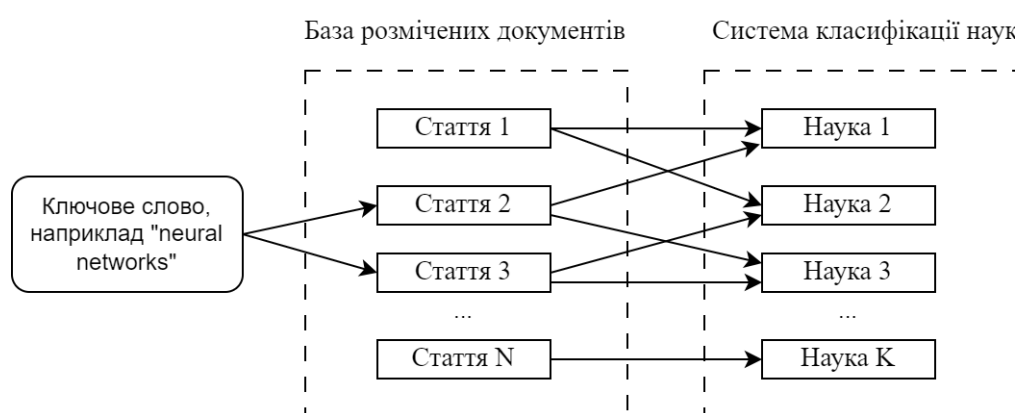


Рисунок 2.3 – Схематичне зображення категоризації ключового слова. Класи ключового слова визначаються по класам статей в яких це ключове слово згадано

2.2.2 Обґрунтування вибору системи класифікації наук та бази наукових публікацій

Для категоризації ключових слів необхідна система наук. Таких систем є багато. Але при виборі враховуємо не лише систему класифікації, але і доступність бази наукових публікацій та відповідних пошукових сервісів. Нам необхідна база наукових публікацій з сервісами автоматичної категоризації публікацій за деякою класифікацією наук. І при цьому ця інформація повинна бути у публічному доступі. Серед таких систем найбільш популярними вважаються Scopus, Springer та Dimensions. Вони надають інтерфейс для

пошуку документів впорядкованих за галуззю/підгалуззю науки за ключовим словом.

Уся наука в Scopus розбита на 27 галузей (табл. 2.1). За такої малої кількості галузей, виокремлення деяких є сумнівним, наприклад, хімічної інженерії.

Таблиця 2.1 – Галузі науки Scopus

№	Галузь науки Scopus
1	Agricultural and Biological Sciences
2	Arts and Humanities
3	Biochemistry Genetics and Molecular Biology
4	Business Management and Accounting
5	Chemical Engineering
6	Chemistry
7	Computer Science
8	Decision Sciences
9	Dentistry
10	Earth and Planetary Sciences
11	Economics Econometrics and Finance
12	Energy
13	Engineering
14	Environmental Science
15	Health Professions
16	Immunology and Microbiology
17	Materials Science
18	Mathematics
19	Medicine
20	Multidisciplinary
21	Neuroscience
22	Nursing
23	Pharmacology Toxico-logy and Pharmaceutics
24	Physics and Astronomy
25	Psychology
26	Social Sciences
27	Veterinary

Springer – міжнародне видавництво, що займається публікацією книг, статей з журналів. Портфель видавництва складає близько 3 тис. журналів та 250 тис. книг. В Springer наука розбита на 24 галузі (рис. 2.4). Окрім галузей існує класифікація за підгалуззями. Їх налічують близько 2 тис. З метою врахування лише значних підгалуззей під час категоризації можна встановити поріг входження підгалузі. Типовий пошук кількості публікацій за ключовим

словом подано на рис. 2.5. Результати пошуку знаходяться у блоках «Discipline», «Subdiscipline». В Springer наукові галузі погано впорядковані – містять як синоніми, так і занадто дрібні екземпляри. Більш цікавим є підгалузі (subdisciplines) Springer, які дозволяють достатньо вузько категоризувати дослідження. Усього таких підгалузей майже 1600. Як і у випадку галузей, деякі підгалузі мають крихітне наповнення. Є також і синонімія. Нами встановлено, що 95% наповнення припадає на 595 підгалузей.



- » Biomedicine
- » Business and Management
- » Chemistry
- » Computer Science
- » Earth Sciences
- » Economics
- » Education
- » Engineering
- » Environment
- » Geography
- » History
- » Law
- » Life Sciences
- » Literature
- » Materials Science
- » Mathematics
- » Medicine & Public Health
- » Pharmacy
- » Philosophy
- » Physics
- » Political Science and International Relations
- » Psychology
- » Social Sciences
- » Statistics

Рисунок 2.4 – Галузі науки у Springer

Dimensions містить у собі статі журналів та показники цитувань. Dimensions індексує понад 140 млн. публікацій. Наука поділяється на 22 галузі та 171 спеціальностей науки. Класифікація наук визначена у системі ANZSRC-2020 (Australian and New Zealand Standard Research Classification), що сформована Австралійським Бюро Статистики (Australian Bureau of Statistics) для оцінювання і аналізу наукових досліджень проведених у Австралії та Новій Зеландії.

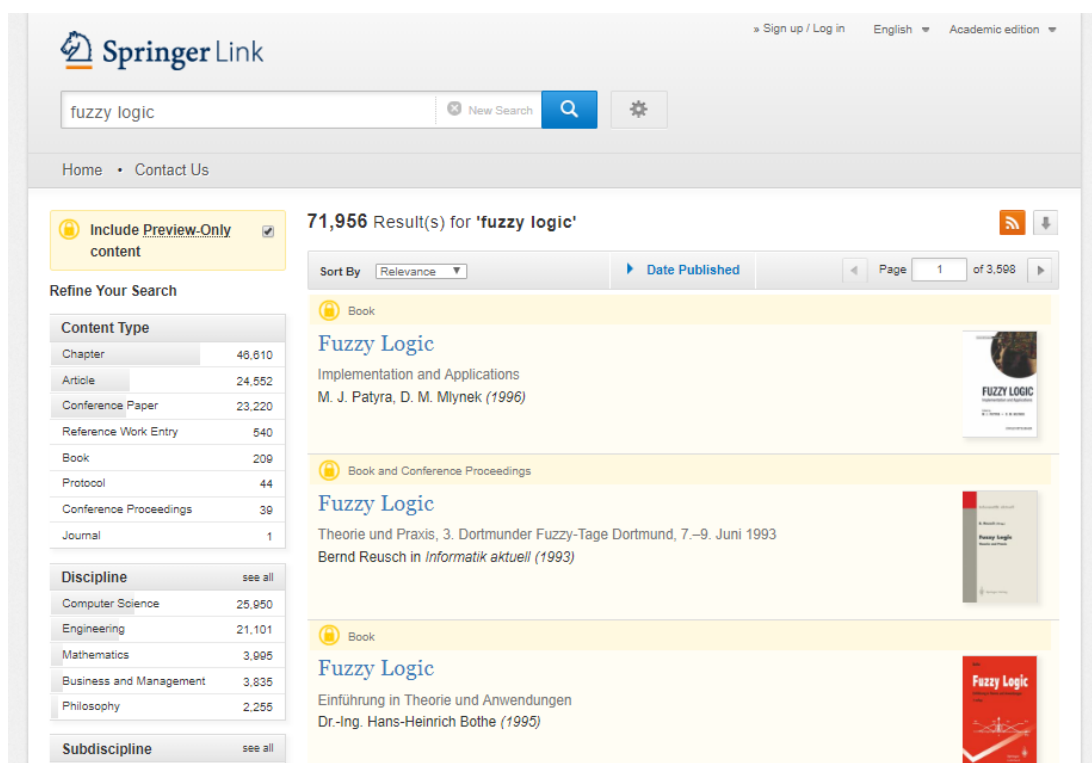


Рисунок 2.5 – Пошук кількості публікацій за ключовим словом у Springer

ANZSRC-2020 – це тривірнева класифікація наук за схемою: галузі (Divisions), спеціальності (Groups) та області (Fields).

Уся наука у Dimensions розділена на такі 22 галузі:

- 30 Agricultural, Veterinary and Food Sciences;
- 31 Biological Sciences;
- 32 Biomedical and Clinical Sciences;
- 33 Built Environment and Design;
- 35 Commerce, Management, Tourism and Services;
- 34 Chemical Sciences;
- 36 Creative Arts and Writing;
- 37 Earth Sciences;
- 38 Economics;
- 39 Education;
- 40 Engineering;
- 41 Environmental Sciences;
- 47 Language, Communication and Cul-ture;

- 48 Law and Legal Studies;
- 42 Health Sciences;
- 43 History, Heritage and Archaeology;
- 44 Human Society;
- 46 Information and Computing Sciences;
- 49 Mathematical Sciences;
- 50 Philosophy and Religious Studies;
- 51 Physical Sciences;
- 52 Psychology.

Кожна галузь об'єднує від 3 до 19 спеціальностей. Назва спеціальності складається з цифрового коду та змістовної частини. Код складається з номеру галузі та порядкового номеру спеціальності в межах галузі. Всього є 171 спеціальність, список яких наведено в Додатку Б.

Типовий приклад пошуку кількості публікацій за ключовим словом подано на рис. 2.6. Запит до Dimensions формуємо окремо за кожним ключовим словом окремо. Якщо цей елемент є словосполученням, тоді подамо його у лапках. Областю пошуку визначимо назву та реферат публікації (Title and Abstract). За кожною спеціальністю та за кожною галуззю виводиться кількість публікацій, в назві або в рефераті яких фігурує пошуковий вираз. Видача відсортована за спаданням кількості публікацій. Також можна отримати загальну кількість публікацій за кожною спеціальністю, тобто обсяги тематичних колекцій. Результат пошуку знаходиться у блоці «Fields of Research». Dimensions індексує переважно англomовні публікації, тому усі інтереси з профіля кандидата доцільно попередньо перекласти на англійську.

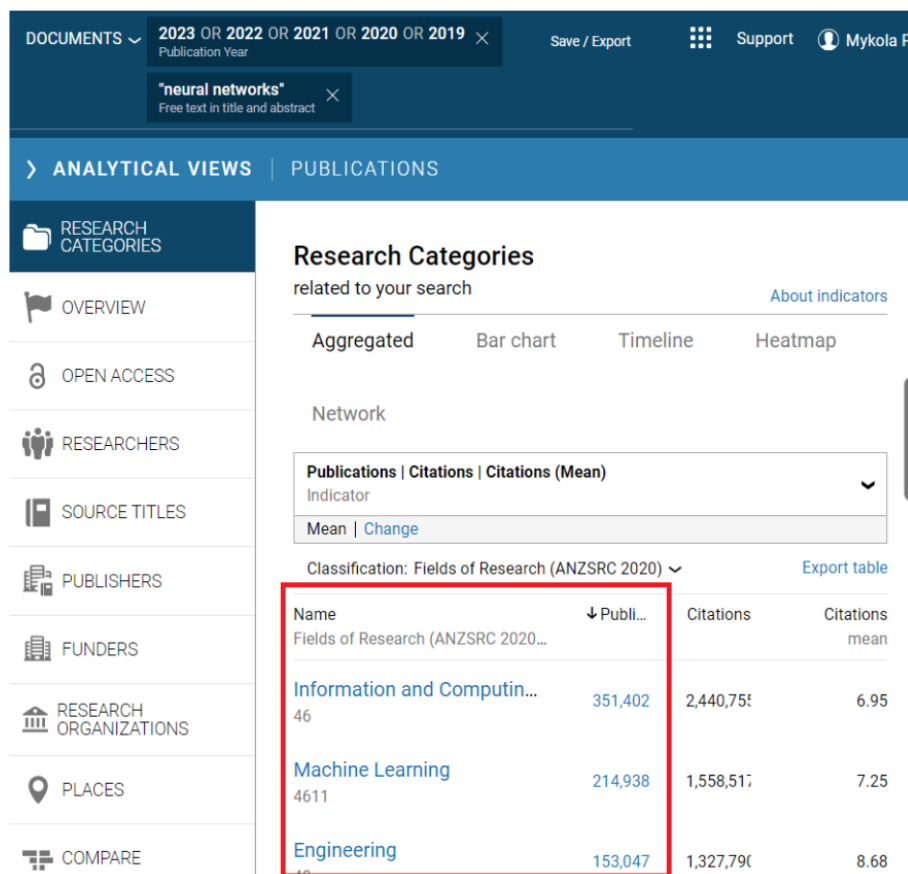


Рисунок 2.6 – Пошук публікацій за ключовим словом у науковій базі Dimensions

На відміну від своїх аналогів, система Dimensions не має вказаних недоліків. Окрім цього, вона безкоштовна та дозволяє отримати кількість спільних документів для двох спеціальностей. Тому для реалізації категоризації використовуватимемо саме цю систему.

Висновки до розділу 2

1. Виконано високорівневий опис пропонованого методу підбору рецензентів. Визначено дані, що використовуються для прийняття рішень.
2. Виконано високорівневий опис представлення даних про заявку та рецензентів як задачу категоризації в межах деякої системи класифікації наук.
3. Встановлено, що системи класифікації наук в інформаційних системах Scopus та Springer мають певні недоліки. У Scopus мала кількість галузей,

виокремлення деяких є сумнівним, наприклад, хімічної інженерії. У Springer як і у випадку галузей, деякі підгалузі мають крихітне наповнення. Є також і синонімія. Нами встановлено, що 95% наповнення припадає на 595 підгалузей. Тому ці інформаційні системи не підходять для проведення дослідження.

4. Встановлено, що найбільш відповідає вимогам дослідження інформаційна система Dimensions, яка використовує дворівневий варіант системи – систему класифікації наук ANZSRC-2020.

Результати розділу опубліковано в таких роботах автора: [13], [14].

РОЗДІЛ 3

СХОЖІСТЬ РЕЦЕНЗЕНТА ТА ТВОРУ З УРАХУВАННЯМ СПОРІДНЕНOSTІ НАУКОВИХ СПЕЦІАЛЬНОСТЕЙ

3.1 Критерій схожості категоріальних розподілів з врахуванням схожості категорій

Для співставлення рецензентів та заявок потрібно знати наскільки схожі два категоріальні розподіли – розподіл інтересів рецензента за спеціальностями та розподіл заявки за спеціальностями. Для цього потрібно визначити критерій схожості двох розподілів. В залежності від типу опису об'єктів використовують різні критерії схожості об'єктів. Для об'єктів у метричному просторі схожість визначають як величину обернену чи інверсну до відстані між двома точками. Координатами кожної точки є числові значення атрибутів відповідного об'єкту. Чим менше відстань між аналізованими об'єктами, тим вони більш схожі. В статті [158] проаналізовано майже 50 різних критеріїв, найбільш популярними серед них є частинні випадки метрики Мінковського – евклідова відстань, манхетенська відстань та метрика Чебишева. Часто використовується також і косинусна метрика, за якою розраховується косинус кута між двома векторами, які виходять з початку координат та прямують до аналізованих об'єктів.

У категоріальному просторі схожість двох об'єктів визначається, зазвичай, як суперпозиція схожості об'єктів за кожною категорією. Найчастіше – це сума схожості за окремими категоріями. При цьому, кожна категорія розглядається незалежно та ізольовано від інших. Є і зворотній підхід, коли спочатку визначають розбіжність об'єктів за кожною категорією, а потім їх агрегують, щоб розрахувати загальну схожість. Один із популярних варіантів такої метрики запропоновано в [159] для розрахунку схожості нечітких множин. В тій статті розбіжність об'єктів визначається через модуль різниці ступенів належності. Усі метрики з оглядової статі [158] та з інших релевантних публікаціях, наприклад, [160], [161] передбачають відсутність

спорідненості між категоріями. Але, для деяких практичних задач категорії є спорідненими. Це призводить до того, що схожість між об'єктами слід розраховувати не лише напрямую, як схожість між еквівалентними категоріями, але і враховувати непряму, перехресну схожість через споріднені категорії. Враховуючи специфіку предметної області в даній роботі ми пропонуємо критерій схожості, який додатково враховує схожість наукових об'єктів через спорідненні категорії.

Розглянемо такий приклад. Нехай рецензента А віднесено до спеціальності «*Системний аналіз*» зі ступенем належності 0.4 та до спеціальності «*Інформаційні системи та технології*» зі ступенем 0.6. Рецензента В віднесено до спеціальності «*Системний аналіз*» зі ступенем належності 0.7 та до спеціальності «*Комп'ютерні науки*» зі ступенем 0.3. Рецензента С віднесено до спеціальності «*Системний аналіз*» зі ступенем належності 0.4 та до спеціальності «*Маркетинг*» зі ступенем 0.6. За будь-якою з відомих метрик схожість між парою наведених вище рецензентів буде встановлено лише за їх належностями до спільної спеціальності «*Системний аналіз*». Належності до інших категорій не враховуються тому, що вони у науковців не співпадають. Схожість між рецензентами А та В визначається виключно на основі їх ступенів належності до категорії «*Системний аналіз*», які дорівнюють 0.4 та 0.7. Якщо схожість визначати за спільною часткою належності, використовуючи операцію мінімуму, отримуємо, що схожість науковців А та В дорівнює $Fit(A,B) = \min(0.4, 0.7) = 0.4$. Аналогічно, схожість рецензентів А та С дорівнює $Fit(A,C) = \min(0.4, 0.4) = 0.4$, а рецензентів В та С дорівнює $Fit(B,C) = \min(0.7, 0.4) = 0.4$. Але, предметна область спеціальностей така, що «*Інформаційні системи та технології*» значно ближче до «*Комп'ютерних наук*», ніж до «*Маркетингу*». Також, «*Комп'ютерні науки*» значно ближче до «*Інформаційні системи та технології*», ніж до «*Маркетингу*». Відповідно, схожість рецензентів А та В має бути вищою, ніж схожість рецензентів А та С чи рецензентів В та С. Але відомі критерії схожості не враховують спорідненість категорій, тому за ними неможливо врахувати такі особливості.

Позначимо кількість категорій через m . Тоді, об'єкти X та Y , схожість яких будемо оцінювати, опишемо такими розподілами належностей до категорій: $(\mu_1(X), \mu_2(X), \dots, \mu_m(X))$ та $(\mu_1(Y), \mu_2(Y), \dots, \mu_m(Y))$. Розподіли вважатимемо нормалізованими, що задовольняють такі умови:

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

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

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

Задача полягає в тому, щоб для об'єктів X та Y розрахувати показник схожості. Специфіка предмету дослідження полягає в тому, що деякі категорії є спорідненими. Відповідно, слід враховувати не лише схожість за ідентичними категоріями, але і за спорідненими. Нижче пропонується такий критерій, який враховує семантичну спорідненість категорій.

Схожість двох об'єктів X та Y пропонується визначити таким чином:

$$Fit(X, Y) = F(X, Y) + \Delta F(X, Y), \quad (3.1)$$

де $F(X, Y)$ – доданок, що оцінює безпосередню (пряму) схожість об'єктів X та Y за категоріями;

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

Перший доданок в формулі (3.1) розрахуємо за спрощеним варіантом метрики Чекановського для випадку, коли ступені належності знаходяться у інтервалі $[0,1]$ і розподіли пронормовані. Розрахункова формула є такою:

$$F(X, Y) = \sum_{i=\overline{1, m}} \min(\mu_i(X), \mu_i(Y)), \quad (3.2)$$

де $\mu_i(X)$ – ступінь належності об'єкта X до i -ї категорії, $i = \overline{1, m}$;

$\mu_i(Y)$ – ступінь належності об'єкта Y до i -ї категорії, $i = \overline{1, m}$.

Формулу (3.2) можна інтерпретувати як суму належностей перетину нечітких множин X та Y . В формулі (3.2) вважається, що загальна схожість двох об'єктів є сумою їх схожостей за кожною категорією. Схожість за

категорією визначається як мінімум належностей обох об'єктів до цієї категорії. Таким чином, у критерій схожості (3.2) один із об'єктів вносить усе значення ступеня належності до категорії, а у другий – лише частину.

Після застосування формули (3.2) отримуємо такі залишки належності:

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

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 вище 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 у критерій (3.1) через попарну спорідненість категорій розрахуємо так:

$$\Delta F(X, Y) = \sum_{i=1, m} \sum_{j=1, m} e_{ij} \cdot k_{ij}. \quad (3.3)$$

Приклад. Задано 2 об'єкти з такими належностями до категорій $\{A, B, C, D\}$: $X = (0.5 \ 0.2 \ 0.1 \ 0.2)$ та $Y = (0.7 \ 0.1 \ 0.2 \ 0)$. Спорідненість

категорій описана такою матрицею: $\mathbf{K} = \begin{pmatrix} 1.0 & 0.5 & 0.0 & 0.0 \\ 0.5 & 1.0 & 0.1 & 0.0 \\ 0.0 & 0.1 & 1.0 & 0.3 \\ 0.0 & 0.0 & 0.3 & 1.0 \end{pmatrix}$. Розрахуємо

схожість об'єктів X та Y за запропонованим критерієм (3.1).

Для розрахунку першого доданку критерію схожості (3.1) зробимо перетин двох розподілів (рис.3.1). Числове значення першого доданку є таким:

$$F(X, Y) = \min(0.5, 0.7) + \min(0.2, 0.1) + \min(0.1, 0.2) + \min(0.2, 0) = 0.5 + 0.1 + 0.1 + 0 = 0.7$$

Залишки після перетину становлять: $e(X) = (0 \quad 0.1 \quad 0 \quad 0.2)$ та

$$e(Y) = (0.2 \quad 0 \quad 0.1 \quad 0). \text{ Композиція залишків дорівнює } \mathbf{E} = \begin{pmatrix} 0.0 & 0.0 & 0.0 & 0.0 \\ 0.1 & 0.0 & 0.1 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.2 & 0.0 & 0.1 & 0.0 \end{pmatrix}.$$

Виконавши поелементний добуток матриць $\mathbf{E}$ та $\mathbf{K}$, отримуємо таку матрицю

внесків через споріднені категорії: $\begin{pmatrix} 0.0 & 0.0 & 0.0 & 0.0 \\ 0.05 & 0.0 & 0.01 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.03 & 0.0 \end{pmatrix}$. З цієї матриці видно,

що внесок від врахування спорідненості другої та першої категорій становить 0.05, внесок від врахування спорідненості другої та третьої категорій становить 0.01, а внесок від врахування спорідненості четвертої та третьої категорій становить 0.03. Внесок через спорідненість інших категорій є нульовим. Сумарний внесок від усіх споріднених категорій становить: $\Delta F(X, Y) = 0.05 + 0.01 + 0.03 = 0.09$. Результуюче значення схожості об'єктів X та Y за формулою (3.1) дорівнює $F(X, Y) = 0.7 + 0.09 = 0.79$.

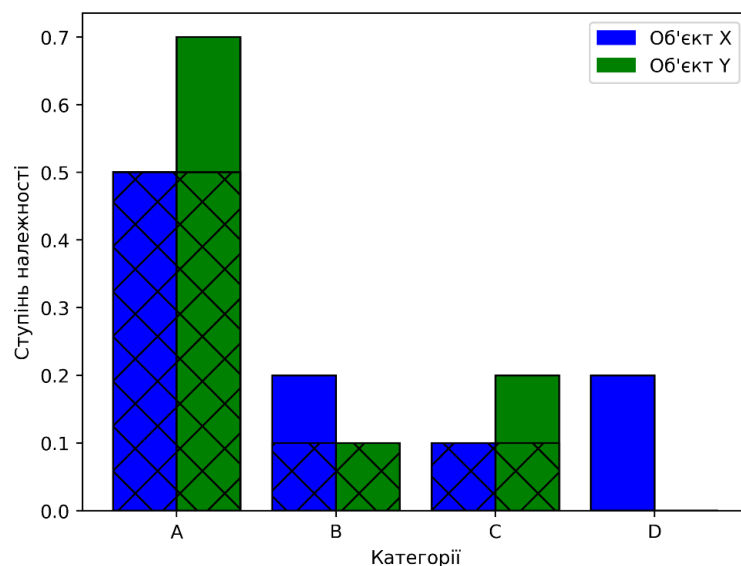


Рисунок 3.1 – Перетин двох категоріальних розподілів для розрахунку $Fit(X, Y)$

3.2 Експериментальні дослідження запропонованого критерію

В наведеному вище прикладі врахування спорідненості категорій збільшило рівень схожості на 0.1, що становить 14% від початкового значення, отриманого за метрикою Чекановського. Проведемо обчислювальні експерименти, щоб встановити наскільки чутливий запропонований критерій до врахування спорідненості категорій.

Виконаємо 20 серій експериментів. В кожній серії досліджуються розподіли з однаковою кількістю категорій m . Кількість категорій від серії до серії збільшується від 3 до 60 з кроком 3. В кожній серії розраховується схожість 3000 пар об'єктів X та Y . Належності об'єктів X і Y та матрицю спорідненості категорій згенеруємо випадково з використанням узагальненого розподілу Парето. Вибір цього розподілу обумовлено тим, що на практиці, зокрема при порівнянні спеціальностей науки, розподіл схожості категорій є подібними до паретівського (див., наприклад, [3], [12]). Для кожного експерименту спочатку випадковим чином згенеруємо параметри розподілів. Параметр форми k згенеруємо з інтервалу $[0.15, 0.5]$, параметр масштабу σ згенеруємо з інтервалу $[0.001, 0.01]$, зміщення приймемо нульовим – $\theta = 0$. У кожному експерименті за синтезованими розподілами випадковим чином згенеруємо координати векторів X і Y та матрицю спорідненості категорій K . В матриці K максимальне значення елементів обмежимо на рівні 0.4; вектори X і Y нормалізуємо на 1. Далі розраховуємо схожість між X та Y за запропонованим критерієм (3.1).

Результати експериментів у формі коробчатих діаграм наведено на рис. 3.2. З рисунка видно, що значення схожості потрапляють в діапазон $[0, 1]$. При збільшенні категорій розкид результатів знижується. Медіана розподілів за $m > 3$ не залежить від кількості категорій.

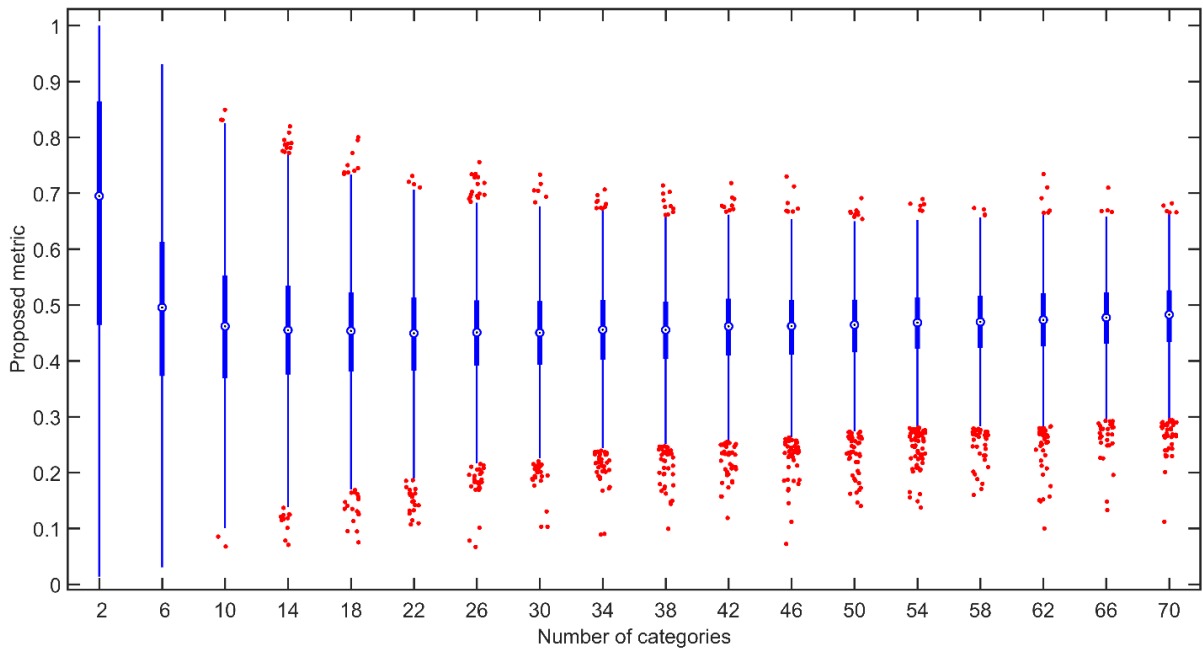


Рисунок 3.2 – Розподіли схожості за запропонованим критерієм

На рис. 3.3 наведено коробчасті діаграми розподілу доданку $\Delta F(X, Y)$, що враховує схожість об'єктів X та Y через споріднені категорії. Медіана цієї величини збільшується від 0 до 0.06 зі збільшенням кількості категорій. Зростає також і розкид розподілів, при цьому кількість викидів спадає. Зростання медіани може свідчити про те, що зі збільшенням категорій значний внесок у $\Delta F(X, Y)$ вносить велика кількість дрібних зв'язків між спорідненими категоріями. Зі збільшенням категорій кількість пар споріднених категорій зростає квадратично. Внесок багатьох пар споріднених категорій є крихітних – шумовим, але сума їх внесків виявляється великою. Зменшення кількості викидів також є негативним чинником. Необхідність нового критерію оцінки схожості обумовлена в першу чергу потребою виявлення специфічних випадків, з сильними перехресним впливом через споріднені категорії. А зменшення кількості викидів свідчить, що шумові спорідненості ускладнюють виявлення таких пар об'єктів.

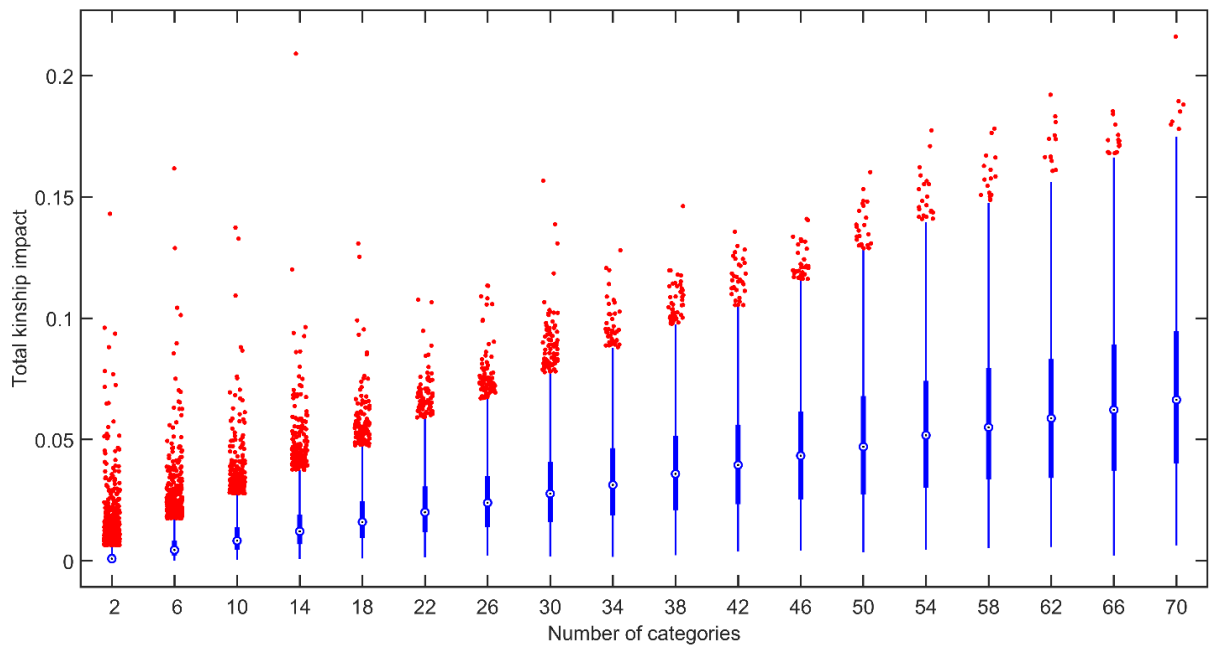


Рисунок 3.3 – Розподіли внеску споріднених категорій

Відфільтруємо шумовий внесок споріднених категорій. Вважатимемо, спорідненість шумовою, якщо коефіцієнт спорідненості менший за 0.05. Коробчасті діаграми розподілу внеску шумової спорідненості наведено на рис. 3.4. З нього видно, що медіана шумового внеску лінійно зростає і досягає значення 0.045 за 60-ти категорій. Щодо відносного внеску шуму (рис. 3.5), то його медіана перевищує 10% у серіях експериментів за великої кількості категорій – 57 та 60. В кожній серії експериментів мають місце численні випадки, коли шумовий внесок перевищує 15%. Шумовий внесок перевищує 30% лише у 5 випадках із 60000.

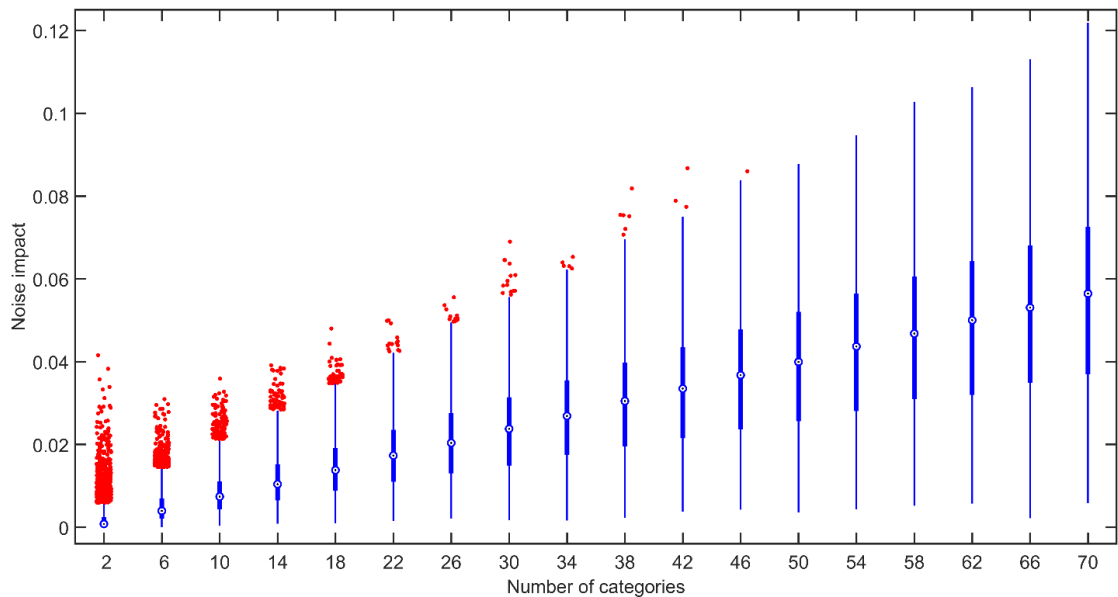


Рисунок 3.4 – Розподіли шумового внеску від врахування споріднених категорій

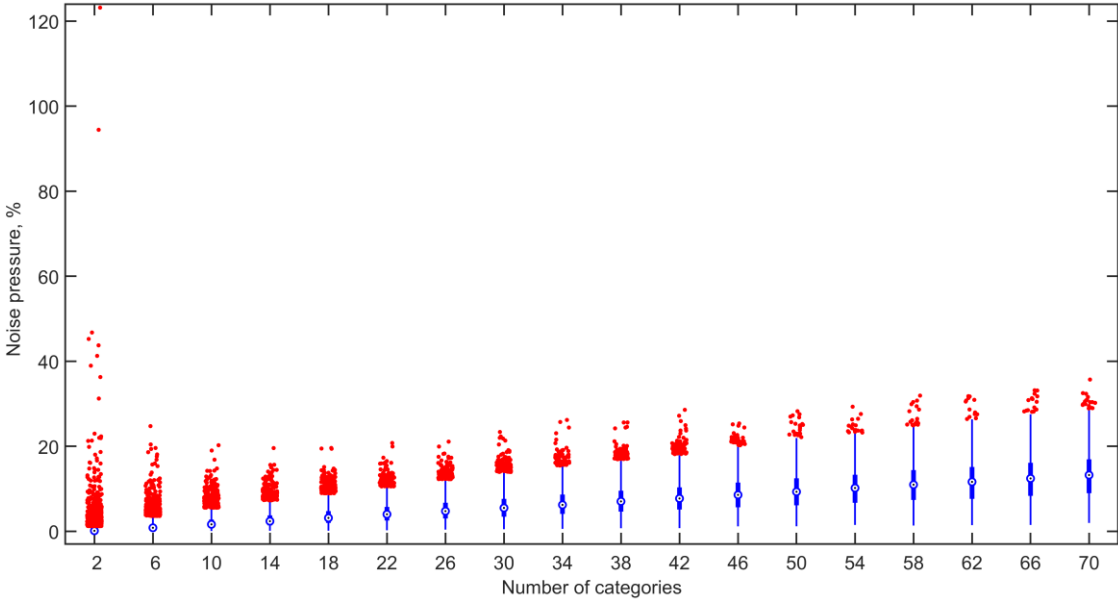


Рисунок 3.5 – Розподіли відносної величини шумового внеску від споріднених категорій

Після видалення шумової складової коробчасті діаграми розподілу внеску споріднених категорій наведено на рис. 3.6. Медіана цієї величини зі збільшенням кількості категорій збільшується лише з 0 до 0.01. Це в 6 разів менше, ніж без видалення шуму. При цьому спостерігається досить велика кількість викидів, що вказує на те, що критерій дозволить ідентифікувати випадки сильної взаємодії через споріднені категорії.

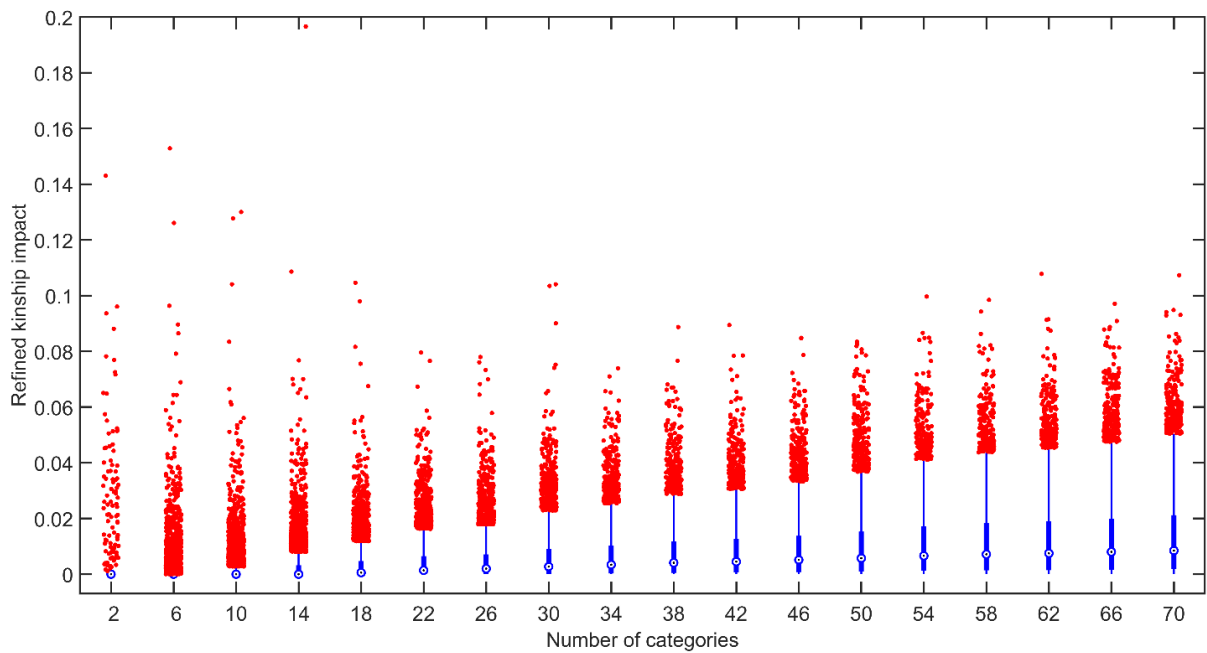


Рисунок 3.6 – Розподіли внеску споріднених категорій після фільтрації шуму

Обчислювальні експерименти показали, що у випадку паретовських розподілів належностей об'єктів за категоріями та паретовських розподілів коефіцієнтів спорідненості категорій, запропонований критерій приймає значення з інтервалу $[0,1]$. Встановлено, що зі збільшенням кількості категорій сильно зростає внесок у запропонований критерій доданку, який враховує спорідненість категорій. Це обумовлено тим, що квадратично зростає кількість дрібних зв'язків між спорідненими категоріями, кожен із яких додає деякий внесок у значення критерію. І хоча внесок від багатьох пар споріднених категорій є крихітним – шумовим, але сума внесків виявляється великою. Для усунення цього недоліку запропоновано не враховувати шумову спорідненість категорій. Простий фільтр з пороговим обмеженням коефіцієнту спорідненості на рівні 0.05 усунув цей недолік. Після такої фільтрації шуму розподіли другого доданку критерію, який враховує спорідненість категорій, мають значну кількість викидів. Значна кількість викидів спостерігається в усіх серіях експериментів, як з малою кількістю категорій, так і з великою. Цей факт вказує на те, що запропонований критерій дозволяє легко ідентифікувати пари

об'єктів, схожість яких визначається значною мірою через належність до споріднених категорій.

В даній роботі ми використовуємо запропонований критерій для порівняння двох розподілів над деякою системою класифікації наук, це можуть бути розподіли ключових слів заявки та рецензента, а також розподіли ключових слів рецензента та розподіл його статей.

3.3 Визначення рівня спорідненості наукових спеціальностей

Запропонований в попередньому підрозділі критерій вимагає інформацію про спорідненість наукових спеціальностей. Найбільш використовуваним підходом для ідентифікації спорідненості вважається аналіз цитування [162]. Проте, останнім часом також почали використовувати методи статистичного аналізу тексту, зокрема в [163] автори пропонують лінгвістичний підхід для дослідження організації та еволюції наукових галузей, що використовуються в Web of Science. В роботі порівнюються три підходи до визначення спорідненості спеціальностей: на основі експертної класифікації; на основі цитувань; на основі лінгвістичної схожості. Оцінити спорідненість пропонувалось метрикою розбіжності (dissimilarity): чим менше значення, тим більша спорідненість, чим більше значення тим менша спорідненість. Експерименти проводились для статей з Web of Science, що розмічені за трьохрівневою системою класифікації наук. Ідея оцінювання спорідненості за цитованістю полягає в тому, що спорідненість між спеціальностями збільшується, якщо в статті однієї спеціальності є посилання на статтю з іншої спеціальності. У випадку лінгвістичної схожості спорідненість визначається за порівнянням частот появи слів в контексті конкретних спеціальностей. Виявлено, що підходи на основі цитувань та лінгвістичного аналізу дають подібні оцінки спорідненості, при цьому значно відрізняються від експертної оцінки спорідненості. Експертна оцінка дає ідеалізоване уявлення про спорідненість спеціальностей, в той час як аналіз цитувань дозволяє виявити

соціальну спорідненість спеціальностей, а лінгвістичний підхід дозволяє виявити змістовну (когнітивну) спорідненість. Приміром, в [164] рівень міждисциплінарності ідентифікують за кількістю цитувань з статті певного дослідника на статті у журналах з Web of Science, де кожен із журналів віднесено до однієї із предметних областей (subject category) Web of Science. В [165] ідентифікація здійснюється за допомогою показників різноманітності за розподілом посилань між дисциплінами з врахуванням відстані між дисциплінами. Дещо менш поширеним підходом є ідентифікація міждисциплінарності за аналізом графу колаборацій над спільними проєктами та співавторствах [166], яка визначається за належністю дослідників до наукових спеціальностей. Як математична модель використовується індекс різноманітності Стірлінга. Деякі дослідження використовують одночасно аналіз цитувань та лінгвістичний аналіз для виявлення рівня міждисциплінарності [167], [168].

Усі вищезгадані підходи вимагають великих інформаційних ресурсів і є трудомісткими. Окрім того, методи, що базуються на аналізі цитування, є досить інерційними. Неможливо миттєво оцінити міждисциплінарність нової роботи, оскільки для її цитування потрібен деякий час. На противагу згаданим підходам нижче пропонується швидка модель оцінювання спорідненості наукових спеціальностями оснований на текстовому аналізі. Модель базується на використанні сервісів інформаційних ресурсів системи Dimensions. Кожна публікація в системі Dimensions віднесена до однієї чи кількох спеціальностей відповідно до системи класифікації наук ANZSRC-2020. Категоризація публікацій в Dimensions здійснена на основі машинного навчання, переважно, на основі аналізу змісту назв, анотацій та ключових слів. Розрахунок спорідненості спеціальностей здійснено за індексом Жаккара як відношення кількості спільних публікацій двох спеціальностей до загальної кількості публікацій за цими двома спеціальностями. До прикладу, на рис. 3.7 подано результат запиту до системи Dimensions за спеціальністю «4609 Information Systems». Як видачу на цей запит система повертає розподіл кількості

публікацій за усіма спеціальностями, які категоризовані до «4609 Information Systems» і одночасно до інших спеціальностей. Наприклад, до спеціальності «4609 Information Systems» та «4605 Data Management and Data Science» одночасно віднесено 22095 публікацій.

За останні 5 років в системі Dimensions проіндексовано 33.8 млн публікацій, які є джерельною базою дослідження. Публікації за спеціальностями розподілені нерівномірно (рис. 3.8). Децильний коефіцієнт дорівнює 14.9. Найбільше публікацій – 3188112 віднесено до спеціальності 3202, найменше – 3339 – до спеціальності 3606. Навіть найменш популярна спеціальність має достатню кількість публікацій, щоб на їх основі отримати достовірні статистичні висновки.

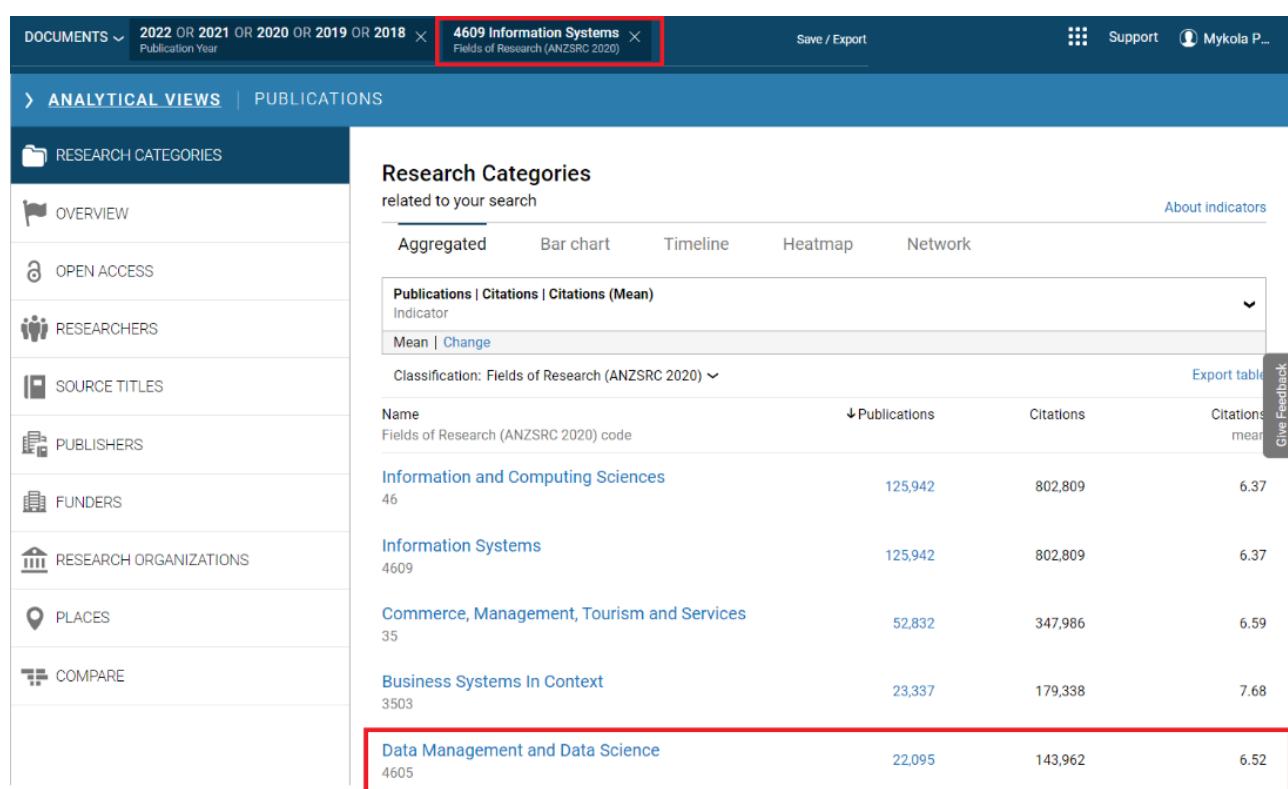


Рисунок 3.7 – Розподіл публікацій, які віднесені до спеціальності «4609 Information Systems», у Dimensions по спеціальностям науки

Для визначення рівня спорідненості між двома спеціальностями введемо наступні позначення:

S_1 – перша спеціальність;

S_2 – друга спеціальність;

k_1 – кількість публікацій віднесених до першої спеціальності;

k_2 – кількість публікацій віднесених до другої спеціальності;

C – кількість публікацій віднесених одночасно до спеціальності 1 та 2.

Тоді рівень спорідненості двох спеціальностей визначимо наступним чином:

$$J(S_1, S_2) = \frac{C}{k_1 + k_2 - C}. \quad (3.4)$$

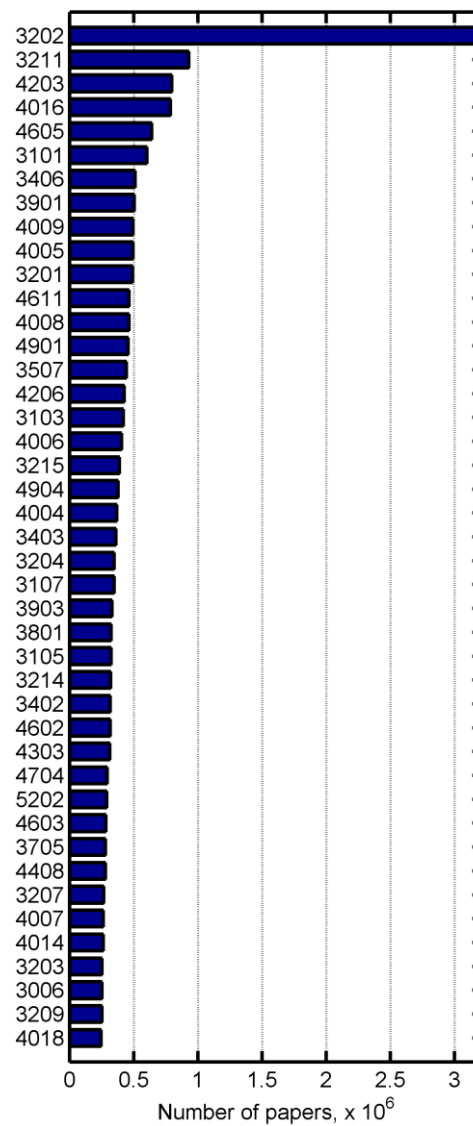


Рисунок 3.8 – Перший квартиль розподілу публікацій за спеціальностями за 2019–2023 рр.

Значення $J(S_1, S_2)$ знаходиться в інтервалі $[0,1]$, де 0 означає відсутність спорідненості, 1 – ідентичність спеціальностей.

Для усіх пар спеціальностей можна сформувати матрицю коефіцієнтів спорідненості:

$$J(S_i, S_j) = \frac{C_{ij}}{k_i + k_j - C_{ij}}, \quad i = \overline{1, m}, \quad j = \overline{1, m} \quad (3.5)$$

де m – кількість спеціальностей;

C_{ij} – кількість публікацій, які відносяться одночасно до i -ї та j -ї спеціальностей;

k_i – кількість публікацій, що віднесені до i -ї спеціальності;

k_j – кількість публікацій, що віднесені до j -ї спеціальності.

Розглянемо приклад визначення спорідненості двох спеціальностей використовуючи систему Dimensions. Розглянемо дві спеціальності «4609 Information Systems» та «4605 Data Management and Data Science». Отримаємо статистику кількості публікацій по кожній спеціальності науки (дані за 2019-2023 роки). Для «4609 Information Systems» та «4605 Data Management and Data Science» відповідно отримуємо кількість публікацій 125942 та 587090, кількість публікацій одночасно віднесених до обох спеціальностей - 22095:

$$\begin{aligned} J(\langle 4609 \text{ Information Systems} \rangle, \langle 4605 \text{ Data Management and Data Science} \rangle) &= \\ &= \frac{22095}{125942 + 587090 - 22095} = 0.032 \end{aligned}$$

Отримати кількість публікації в Dimensions можна за допомогою API [169]. Дані можна отримати використовуючи звичайний HTTP запит, де тілом є текст у форматі Dimensions Search Language [170]. Наприклад, для того щоб отримати розподіл кількості публікацій за кожною спеціальністю за останні 5 років необхідно відправити запит з таким тілом:

```
search publications where year in
[2019,      2020,      2021,      2022,      2023]      return
category_for_2020
```

Надіславши API-запит для кожної спеціальності отримуємо необхідні початкові дані для розрахунку рівня спорідненості усіх пар спеціальностей. Ранговий розподіл пар спеціальностей за індексом Жаккара (3.4) наведено на рис. 3.9. Високу спорідненість має 20 пар спеціальностей, індекс Жаккара для яких перевищує 0.2 (табл. 3.1). Поміж 20 сильно споріднених пар 3 пари утворено спеціальностями з різних галузей знань. Середня спорідненість має місце для 41 пари спеціальностей (табл. 3.2). Поміж 41 пари із середньою спорідненістю 9 пар утворено спеціальностями з різних галузей знань. Спорідненість пар з високими та середніми індексами Жаккара варто врахувати під час автоматичного підбору рецензентів чи виявлення схожих наукових та освітніх установ. Можливо, деякий ефект буде і від врахування пар спеціальностей з низьким рівнем спорідненості. Таких пар виявилось 70. Решта пар спеціальностей, а саме $14535 - 131 = 14404$ має шумову спорідненість – їх індекс Жаккара менше за 0.05. Поріг щодо шуму встановлено на підставі обчислювальних експериментів запропонованого критерію схожості категоріальних розподілів (див. підр. 3.2).

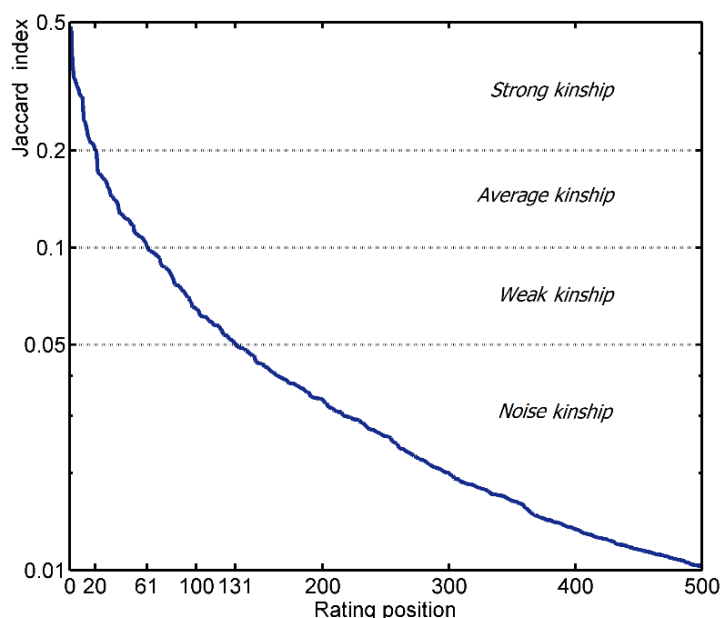


Рисунок 3.9 – Фрагмент рангового розподілу спорідненості пар спеціальностей (напівлогарифмічний формат)

Таблиця 3.1 – Спеціальності з сильною спорідненістю

Спеціальності		Індекс Жаккара
3504	3508	0.485
4703	4704	0.378
3506	3508	0.339
3504	3506	0.332
4008	4009	0.322
4002	4017	0.313
3002	3004	0.309
3901	3903	0.297
5106	5107	0.294
4402	4805	0.293
3503	3506	0.247
4901	4904	0.245
4007	4010	0.236
5106	5110	0.224
3705	3706	0.218
3703	3706	0.212
3703	3705	0.211
4902	5107	0.21
3605	4701	0.208
4004	4011	0.2

Таблиця 3.2 – Спеціальності з середньою спорідненістю

Спеціальності		Індекс Жаккара
3802	3803	0.199
3302	4005	0.172
3502	3801	0.169
4901	4903	0.168
3501	3502	0.167
3801	3802	0.164
5102	5108	0.163
5004	5005	0.162
4902	4904	0.159
3402	3405	0.154
4006	4613	0.152
4203	4205	0.146
4604	4606	0.144
3103	3104	0.144
5201	5205	0.14
4803	4807	0.14
3304	3509	0.139
3102	3105	0.138
3209	5202	0.129
3705	3709	0.127
4407	4408	0.127
5201	5203	0.126
4605	4606	0.124
3801	3803	0.123

4006	4009	0.123
4602	4605	0.122
5101	5109	0.122
3902	3903	0.119
3204	3211	0.118
4804	4807	0.118
3404	3405	0.112
3403	4016	0.111
3406	4016	0.11
5202	5204	0.109
4903	4904	0.108
4803	4804	0.108
3702	3709	0.107
3503	4609	0.106
4301	4303	0.105
5002	5003	0.104
3107	3207	0.101

На рис. 3.10 наведені діаграми спорідненості для 32 спеціальностей, які утворили пари з максимальними індексами Жаккара. Більшість із них мають сильну спорідненість лише з однією спеціальністю. В той же час, 7 спеціальностей сильно взаємодіють з двома спеціальностями. Спеціальність 3506 має високу спорідненість аж з трьома спеціальностями – 3504, 3508 та 3503. Усі вони належать до спільної галузі 35 – Commerce, Management, Tourism and Services. Пари спеціальностей для яких індекс Жаккара більше за 0.05 і які використовуються в даній роботі подано в Додатку Б.

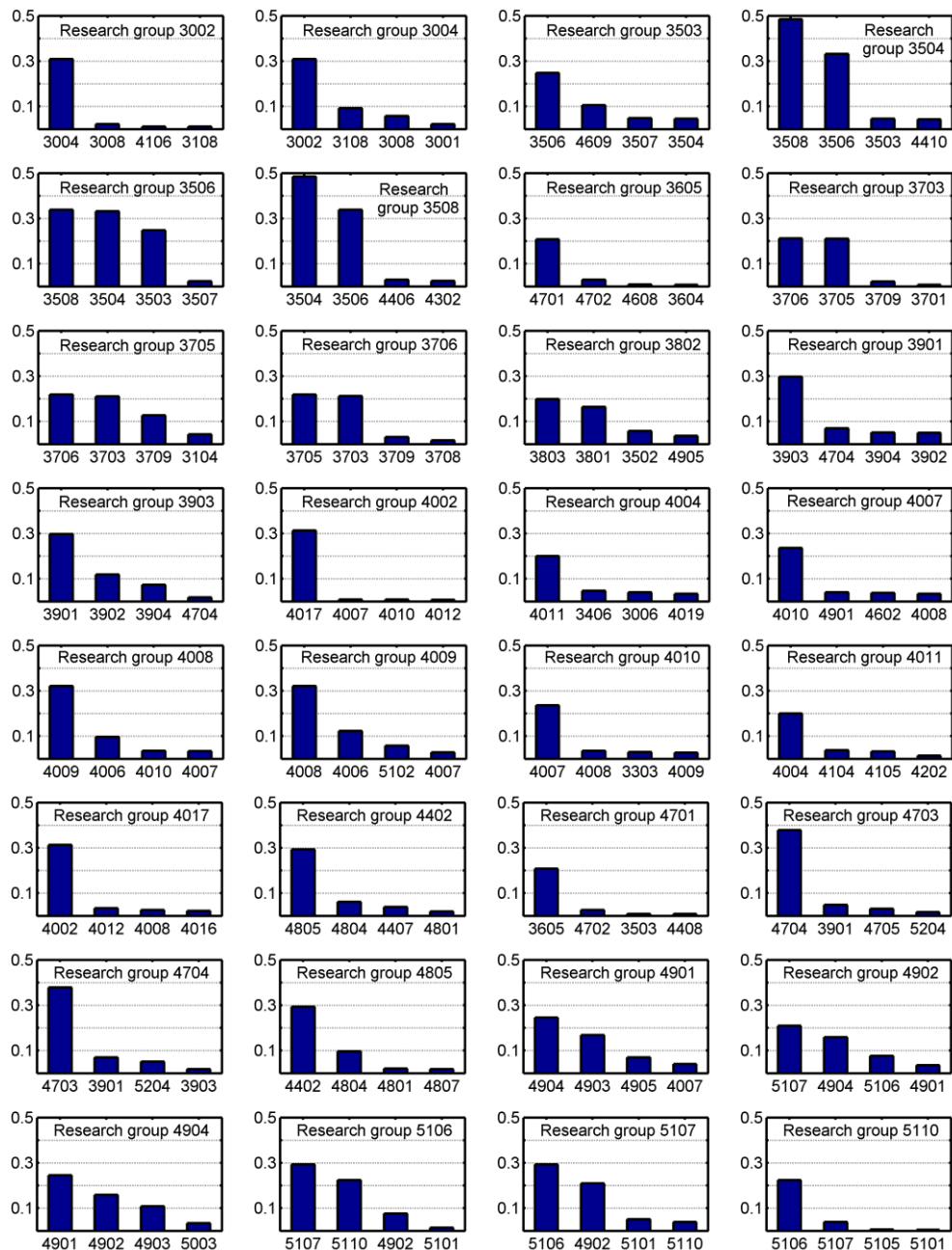


Рисунок 3.10 – Діаграми спорідненості спеціальностей з топ-20 за індексом Жаккара

3.4 Оцінка рівня спорідненості вітчизняних спеціальностей

В Україні діє дворівневою система класифікації освітньо-наукової діяльності на рівні підготовки PhD. Вітчизняні наукові журнали також отримують фахову категорію за спеціальностями з цієї класифікації. Дворівнева система започаткована у 2015 р., і з тих пір зазнавала чотирьох корекцій у 2016, 2017, 2021 та 2022 рр. На сьогодні в Україні є 28 галузей знань. Кожна галузь

знань містить від 1 до 9 спеціальностей. Всього є 121 спеціальність, їх список наведено в Додатку В.

Назва кожної спеціальності складається з цифрового шифру та змістовного словосполучення. Надалі для позначення спеціальності будемо використовувати її цифровий шифр. Перші дві цифри в шифрі відповідають номеру галузі, а остання – порядковому номеру спеціальності в межах галузі.

Напрямку застосувати описані підходи в попередньому підрозділі до оцінювання спорідненості вітчизняних спеціальностей неможливо через такі дві причини. Перша причина – відсутність бази наукових публікацій з відповідною категоризацією кожної публікації до вітчизняних спеціальностей. Друга причина полягає у відсутності відповідності між вітчизняними спеціальностями та спеціальностями ANZSRC-2020. Разом з тим, запропоноване модель можна застосувати для оцінювання спорідненості українських спеціальностей, але не на базі категоризованих публікацій, а на основі інших інформаційних ресурсів. Нами виявлено, що є реєстр експертів НАЗЯВО, кожного із яких віднесено до однієї, двох чи трьох спеціальностей. Експерт НАЗЯВО володіє необхідними знаннями, щоб достовірно оцінити якість освітніх та освітньо-наукових програм за різними спеціальностями. Відповідність експерта НАЗЯВО кільком спеціальностям вважатимемо індикатором їх спорідненості. За такого припущення, можна для ідентифікації рівня спорідненості спеціальностей використовувати запропонований метод, але замість бази наукових публікацій використовувати базу експертів НАЗЯВО.

Базу експертів сформуємо за двома публічними реєстрами НАЗЯВО – «Реєстром експертів з числа науково-педагогічних, наукових працівників» та «Реєстром експертів з числа науково-педагогічних, наукових працівників» чинними на 1 лютого 2024 р. Після об'єднання реєстрів очистимо дані – усі спеціальності приведемо до поточного переліку, наприклад, спеціальність 151 замінимо на 174, 152 – на 175 тощо та видалимо дублікати. У випадку розбіжності цифрового шифру та змістовної назви, спеціальність експерту зазначимо за змістовною назвою у реєстровому записі. В результаті

сформована база включає 4245 експертів, з них 2414 є експертами за однією спеціальністю, 1361 є експертами за двома спеціальностями, і 470 є експертами за трьома спеціальностями. Розподіл експертів за сорока найбільш популярними спеціальностями наведено на рис. 3.11. За спеціальностями 257, 145, 225, 252, 256, 027, 112, 135, 224 експертів мало – від 0 до 5 осіб, тому достовірність статистичних висновків за цими спеціальностями буде низькою. За спеціальностями 143, 173, 187, 021, 194, 203, 207, 026, 134 та 272 кількість експертів складає від 6 до 9, тому використовувати результати статистичних висновків за цими спеціальностями також слід обережно.

Індекс спорідненості пари спеціальностей (A, B) пропонується розрахувати за формулою (3.4), але використовуючи дані розподілу експертів НАЗЯВО. В цьому контексті формула (3.4) має вигляд:

$$J(A, B) = \frac{N_{A \cap B}}{N_A + N_B - N_{A \cap B}}, \quad (3.6)$$

де N_A – кількість експертів за спеціальністю A ;

N_B - кількість експертів за спеціальністю B ;

$N_{A \cap B}$ - кількість осіб, які одночасно є експертами як за спеціальністю A , так і за спеціальністю B .

Формула (3.6) еквівалента (3.4) за виключенням того, що як початкові дані використовується база категоризованих експертів, тоді як в (3.6) база категоризованих наукових публікацій.

Проілюструємо застосування формули (3.6) на прикладі розрахунку спорідненості спеціальностей *113 Прикладна математика* та *124 Системний аналіз*. За спеціальністю 113 в базі є 42 експерта, за спеціальністю 124 – 27 експертів. Експертами одночасно за обома спеціальностями є 6 осіб. Відповідно, рівень спорідненості спеціальностей 113 та 124 становить:

$$J(\langle 113 \text{ Прикладна математика} \rangle, \langle 124 \text{ Системний аналіз} \rangle) = \frac{6}{42 + 27 - 6} = 0.095.$$

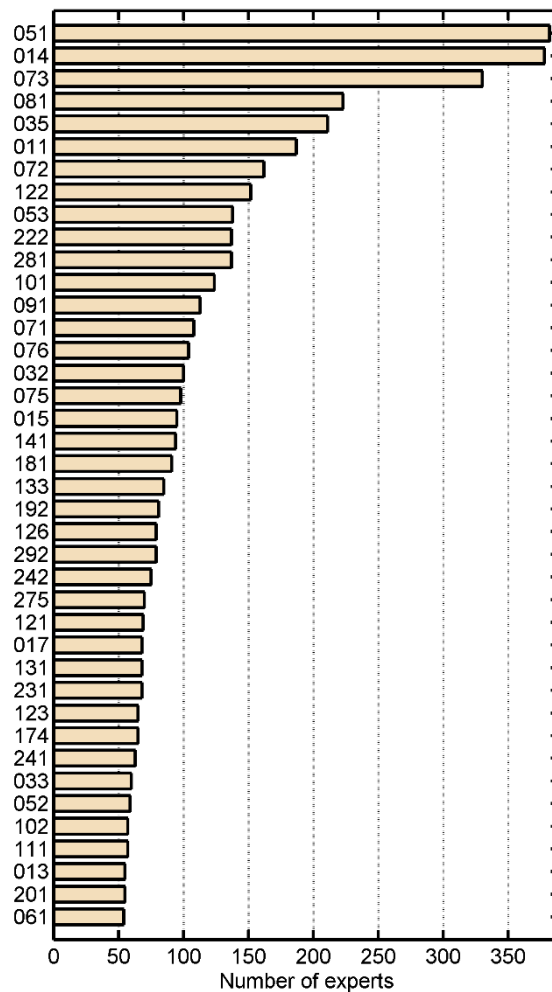


Рисунок 3.11 – Фрагмент розподілу експертів НАЗЯВО за спеціальностями

Ранговий розподіл пар спеціальностей за індексом Жаккара (3.6) наведено на рис. 3.12. Високу спорідненість має 21 пара спеціальностей, індекс Жаккара для яких перевищує 0.1. Для 52 пар спеціальностей спорідненість середня, індекс Жаккара для них знаходиться в інтервалі $[0.05, 0.1]$. Більшість із 7381 пар спеціальностей мають майже нульову спорідненість.

Топ-40 найбільш споріднених пар спеціальностей наведено на рис. 3.13. Для 16 із цих 40 пар спеціальності належать до різних галузей знань. Для 16 із цих 40 пар спеціальності належать до різних галузей знань. Якщо розглядати лише спеціальності з високою спорідненістю, то із 21 пар 9 пар утворюють спеціальності із різних галузей знань. Таким чином, спільна галузева належність не тотожна реальній спорідненості спеціальностей. Список усіх пар

спеціальностей з коефіцієнтом спорідненості більшим за 0.05 наведено в Додатку Г.

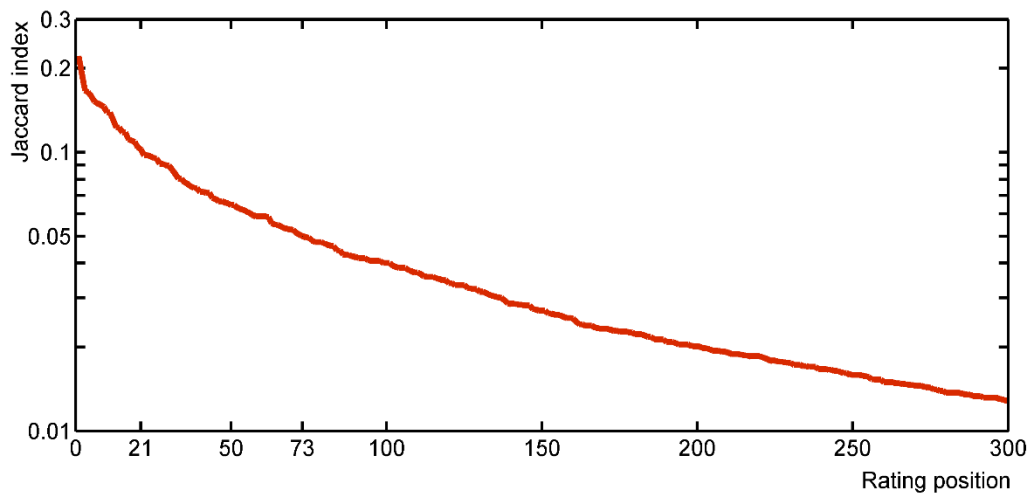


Рисунок 3.12 – Фрагмент рангового розподілу спорідненості пар спеціальностей (напівлогарифмічний формат)

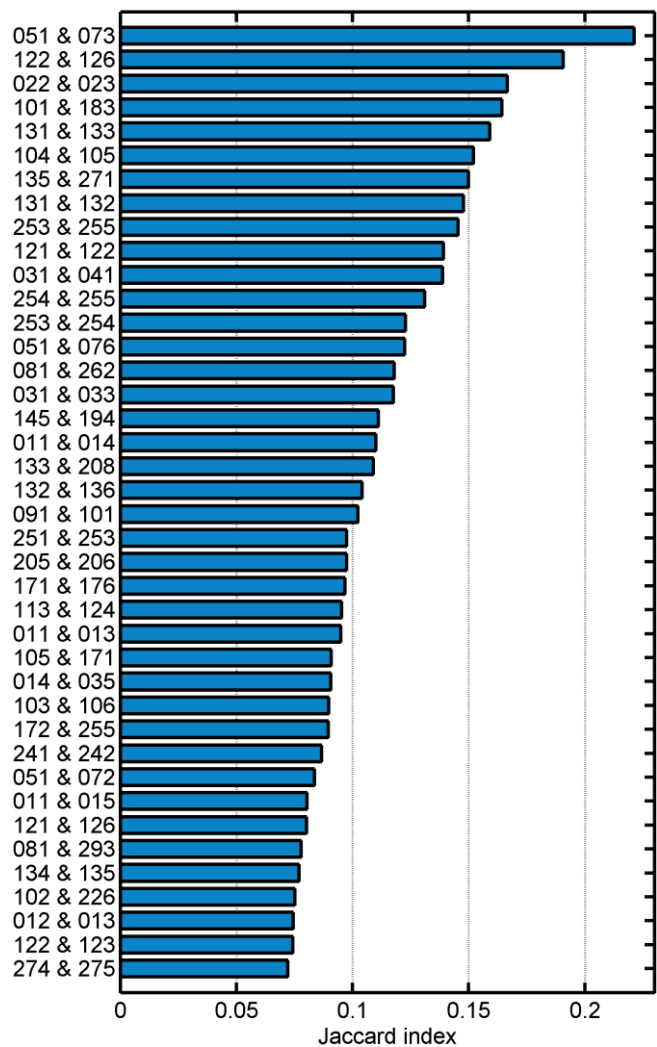


Рисунок 3.13 – Найбільш споріднені пари спеціальностей

Висновки до розділу 3

1. Виконано огляд метрик для порівняння категоріальних розподілів. Встановлено, що вони не враховують спорідненість категорій. Але, для деяких практичних задач категорії є спорідненими. Це призводить до того, що схожість між об'єктами слід розраховувати не лише на пряму, як схожість між еквівалентними категоріями, але і враховувати непряму, перехресну схожість через споріднені категорії.

2. Для врахування спорідненості категорій запропоновано критерій, який удосконалює критерій Чекановського за рахунок додаткового доданку, який враховує перехресну спорідненість категорій, що дозволяє взяти до уваги спорідненість наукових спеціальностей рецензента та заявки.

3. Проведено експерименти над запропонованим критерієм і виявлено, що у випадку паретовських розподілів належностей об'єктів зі збільшенням кількості категорій сильно зростає внесок доданку, що враховує спорідненість категорій. Це обумовлено тим, що квадратично зростає кількість дрібних зв'язків між спорідненими категоріями, кожен із яких додає деякий внесок у значення критерію. І хоча внесок від багатьох пар споріднених категорій є крихітним – шумовим, але сума внесків виявляється великою. Для усунення цього недоліку запропоновано не враховувати шумову спорідненість категорій. Простий фільтр з пороговим обмеженням коефіцієнту спорідненості на рівні 0.05 усунув цей недолік. Після такої фільтрації шуму розподіли другого доданку критерію, який враховує спорідненість категорій, мають значну кількість викидів. Значна кількість викидів спостерігається в усіх серіях експериментів, як з малою кількістю категорій, так і з великою. Цей факт вказує на те, що запропонований критерій дозволяє легко ідентифікувати пари об'єктів, схожість яких визначається значною мірою через належність до споріднених категорій.

4. Проведено аналіз моделей оцінювання спорідненості двох наукових спеціальностей. Встановлено, що вони вимагають великих інформаційних

ресурсів і є трудомісткими. Окрім того, підходи, що базуються на аналізі цитування, є досить інерційними. Неможливо миттєво оцінити міждисциплінарність нової роботи, оскільки для її цитування потрібен деякий час.

5. Вперше розроблена модель статистичного оцінювання спорідненості наукових спеціальностей, що ґрунтується на розрахунку індекса Жаккара за базою категоризованих наукових публікацій, що дозволяє швидко ідентифікувати рівень спорідненості спеціальностей в порівнянні з трудомісткими аналогами на основі аналізу списку цитованості та списку джерел кожної окремої статті

6. Проаналізовано спорідненість спеціальностей з системи класифікації наук ANZSRC-2020. Високу спорідненість має 20 пар спеціальностей, індекс Жаккара для яких перевищує 0.2. Середня спорідненість має місце для 41 пари спеціальностей. Спорідненість пар з високими та середніми індексами Жаккара варто врахувати під час автоматичного підбору рецензентів чи виявлення схожих наукових та освітніх установ. Можливо, деякий ефект буде і від врахування пар спеціальностей з низьким рівнем спорідненості. Таких пар виявилося 70. Решта пар спеціальностей, а саме $14535 - 131 = 14404$ має шумову спорідненість – їх індекс Жаккара менше за 0.05. Поріг щодо шуму встановлено на підставі проведених у розділі обчислювальних експериментів критерію оцінювання схожості категоріальних розподілів.

7. Запропоновану модель спорідненості двох спеціальностей застосовано для оцінювання рівня спорідненості вітчизняних освітніх спеціальностей на основі аналізу профілей експертів НАЗЯВО. Більшість із 7260 пар спеціальностей мають майже нульову спорідненість. Середній рівень спорідненості виявлено у 52 пар спеціальностей, а високу спорідненість має 21 пара. При чому, 9 із 21 пари з високою спорідненістю утворюють спеціальності із різних галузей знань.

Результати розділу опубліковано в таких роботах автора: [3], [4], [5], [6], [7], [8], [9], [10], [18], [19].

РОЗДІЛ 4

РОЗРОБКА МЕТОДУ ЕКСПРЕС-ПІДБОРУ РЕЦЕНЗЕНТІВ

4.1 Формалізована постановка задачі категоризації ключових слів в межах системи класифікації наук

Вважатимемо відомими:

$W = (w_1, w_2, \dots, w_n)$ – список ключових слів, які взято з профілю рецензента або з профілю заявки;

$T = (t_1, t_2, \dots, t_m)$ – перелік наукових спеціальностей у формі списку наукових спеціальностей за деякою класифікацією наук;

$D_1, D_2, \dots, D_m$ – тематичні колекції розмічених текстів, кожна з яких містить лише публікації з спеціальностей $t_1, t_2, \dots, t_m$, відповідно;

$B = D_1 \cup D_2 \cup \dots \cup D_m$ – загальна колекція розмічених текстів, тобто множина публікацій, кожна з яких віднесена до однієї або декількох спеціальностей з множини T ;

$R(D, T) \subset D \times T$ – відношення, яке описує належність публікацій до тематичних колекцій.

Задача полягає у знаходженні тем з T , яким відповідає сукупність ключових слів W . Будемо вказувати не лише сам факт належності, але і ступінь належності. Таким чином на виході отримуємо категоріальний розподіл W на універсальній множині тем T :

$$W^* = (\mu_W(t_1), \mu_W(t_2), \dots, \mu_W(t_m)) \quad (4.1)$$

де $\mu_W(t_p) \in [0, 1]$ – ступінь належності сукупності інтересів W до спеціальності t_p , $p = \overline{1, m}$.

На W^* накладемо такі обмеження:

1) потужність W^* має бути невеликою $1 \leq |W^*| \leq T_{\max}$, наприклад, за $T_{\max} \in \{2, 3, 4\}$ ключові слова відповідатимуть лише кільком спеціальностям;

$$2) \sum_{p=1, m} \mu_W(t_p) = 1.$$

4.2 Категоризація ключових слів до спеціальностей науки з ANZSRC-2020

Категоризацію ключових слів здійснимо на базі таких принципів:

- *принцип статистичної підтримки* – чим більша частка публікацій з певної спеціальності містить аналізоване ключове слово, тим більша належність ключового слова до цієї спеціальності;

- *принцип багатоярликості* – ключове слово може належати до кількох спеціальностей;

- *принцип фільтрації шумів* – ігноруються спеціальності, до яких ключове слово належить з незначним ступенем;

- *принцип ігнорування стоп-слів* – ігнорується ключове слово, яке зустрічається у дуже багатьох категоризованих публікаціях;

- *принцип солідарності* – чим більше ключових слів за окремими запитами віднесено до однієї і тієї ж спеціальності, тим сильніша можливість віднесення науковця до цієї спеціальності;

- *принцип фокусування* – якщо в тематичній колекції багато публікацій, які містять кілька ключових слів одночасно, тоді збільшуються шанси віднесення цих ключових слів до відповідної спеціальності.

- *принцип компактності* – одне ключове слово може бути віднесене лише до невеликої кількості спеціальностей;

- *принцип взаємодії спеціальностей* – під час відсікання хвоста розподілу тем, внесок мінорних спеціальностей перерозподіляється на лідерів з урахуванням їх схожості.

Категоризацію будемо виконувати саме до спеціальностей науки визначених у системі класифікації наук ANZSRC-2020, оскільки кількість

галузей науки в цій системі відносно мала та не дозволить конкретизувати експертну галузь ключових слів.

Наведені принципи пропонується реалізувати алгоритмом, який містить 3 етапи:

1. Етап попередньої обробки ключових слів;
2. Етап пошуку належності кожного ключового слова до спеціальностей науки;
3. Етап узагальнення результатів належності по кожному ключовому слову (принцип солідарності та фокусування).

На першому етапі формується множина пошукових запитів E використовуючи початкові ключові слова та їхні попарні комбінації. На другому етапі обчислюються ступені належності пошукових запитів до спеціальностей. Для цього спочатку знаходиться розподіл кількості публікацій по спеціальностям за останні 5 років використовуючи API Dimensions. Це все виконується для кожного пошукового запиту з відповідним виявленням стоп слів та фільтром шумів. Після цього отримуємо відносну частоту появи пошукових запитів в контексті всіх спеціальностей, та фільтруємо шумові частоти використовуючи параметр *Tail*. На третьому етапі усі розподіли частот по кожному пошуковому запиту усереднюються утворюючи одновимірний вектор. Далі виконуємо обрізання цього вектору до щонайбільше RG_max спеціальностей з ненульовим ступенем належності. Сукупність ключових слів може бути категоризована в кінцевому випадку до щонайбільше $T_max=4$ спеціальностей, і найменший ступінь належності при цьому повинен бути не менше за RG_min_degree . Псевдокод запропонованого алгоритму має наступний вигляд:

```
%RG = Спеціальність науки (в Dimensions - Research Group)
%Алгоритм категоризації
```

```
% #1 - створення множини E пошукових запитів із ключових слів
E=W
```

```
for i=1:length(W)-1
    for j=i:length(W)
        E={E; [W(i) 'AND' W(j)] }
```

```

end
end
% #2 - розрахунок належностей до спеціальностей за кожним запитом
< Знаходимо кількість публікацій по кожній спеціальності N=[N(1),
N(2), ..., N(m)], m=171 >
Counter=0 % лічильник вдалих пошукових запитів
for i=1:length(E)
    < Знаходимо Q - кількість документів в Dimensions, які
    містять E{i} >
    If Q>Threshold_SW continue % ідентифікація стоп-слів
end
    If Q<Threshold_noise continue; % ідентифікація шуму
end
    < Знаходимо t(1), t(2),..., t(m) - кількість документів в
    тематичних колекції за кожною спеціальністю за запитом E{i} >
    % Занулюємо дуже малі видачі:
    index=find(t<Threshold_topic)
    t(index)=0
    if max(t)==0 continue
end
r=t./N % частота входження запиту E{i} до спеціальностей;
% Нормуємо частотні розподіли:
Gamma=r./sum(r)
< Обираємо найбільш популярні спеціальності, які мають
кумулятивний внесок в Gamma не нижче Tail. Спеціальності, що мають
кумулятивний внесок нижче за Tail, знаходяться у векторі Rejected
>
% Відкидаємо спеціальності, що мають кумулятивний внесок нижче за
Tail:
Gamma(Rejected)=0
Gamma=Gamma./sum(Gamma) % Нормуємо знову
Counter=Counter+1
Mu(Counter)=Gamma
end
If Counter==0
    return ('Unsuccessful')
end
% #3 - узагальнюємо належності по кожному ключовому слову
Mu_mean=mean(Mu) % усереднення по усім ключовим словам
% Підрахунок поточної кількості вибраних спеціальностей:
Current_N_RGs=sum(Mu_mean>0);
[Mu, RG_ID, Current_N_RGs]=Top_RG(Mu_mean, Source_RG_ID, RG_max);
% Top_RG - створює RG_ID як вибірку з RG_max спеціальностей з
найбільшим ступенем належності з Source_RG_ID. RG_ID це список
спеціальностей у спадному порядку відповідно до їхніх ступенів
належності Mu. Вектор Mu пронормований у [0;1].
while true
    % RG_min_degree - це найменший дозволений ступінь належності
    до спеціальності
    if (Current_N_RGs < 5 and Mu(end) > RG_min_degree) break
    if (Current_N_RGs <= 1) break
end

```

```

< Виключити спеціальність з низьким рівнем належності і
перерахувати її внесок на інших на основі схожості
спеціальностей>
for target = 1: Current_N_RGs-1
    sim_factor=Jaccard(RG_ID(target), RG_ID(Current_N_RGs));
    Mu(target) = Mu(target) + Mu(Current_N_RGs)*sim_factor;
end
[Mu, RG_ID, Current_N_RGs]=Top_RG(Mu, RG_ID, Current_N_RGs-
1);
end
Return(Mu, RG_ID)

```

На останньому етапі роботи алгоритму під час ітераційного відкидання мінорної спеціальності її внесок перерозподіляється на інші спеціальності з урахуванням рівня спорідненості двох спеціальностей визначених у попередньому розділі. Рівень спорідненості збільшує рівень належності на величину пропорційну до рівня спорідненості між спеціальністю, що відкидається і всіма іншими спеціальностями що залишаються. Рівень спорідненості оцінюється за допомогою індексу Жаккара де розмір перетину це кількість публікацій категоризованих до обох спеціальностей одночасно, і розмір об'єднання це кількість публікацій категоризованих до першої або другої спеціальності. Мотивація цього кроку полягає у тому, що ми збільшуємо вплив підмножини спеціальностей які більш споріднені між собою.

Наприклад, нехай на проміжному етапі ключові слова віднесено до наукових спеціальностей *4410 Sociology*, *4611 Machine Learning*, *3508 Tourism*, та *3504 Commercial Services* таким чином: $W^* = \left(\frac{0.4}{4410}, \frac{0.25}{4611}, \frac{0.2}{3508}, \frac{0.15}{3504} \right)$.

Вилучимо мінорну спеціальність *3504*. Для цього спочатку за запропонованим у попередньому розділі методом визначення спорідненості двох спеціальностей знайдемо коефіцієнти Жаккара між спеціальністю *3504* та іншими. Вони за даними 2019 – 2023 рр. становлять:

$$J(4410, 3504) = 0.044;$$

$$J(4611, 3504) = 0;$$

$$J(3508, 3504) = 0.478.$$

З урахуванням спорідненості внесок мінорної спеціальності 3504 перерозподіляється таким чином:

$$W^* = \left(\frac{0.4 + 0.044 \cdot 0.15}{4410}, \frac{0.25 + 0 \cdot 0.15}{4611}, \frac{0.2 + 0.478 \cdot 0.15}{3508} \right).$$

Порахувавши, отримуємо: $W^* = \left(\frac{0.466}{4410}, \frac{0.25}{4611}, \frac{0.271}{3508} \right)$. Після нормування

до інтервалу $[0,1]$, маємо такий результат: $W^{**} = \left(\frac{0.472}{4410}, \frac{0.253}{4611}, \frac{0.275}{3508} \right)$. В

результаті спеціальність 3508 *Tourism* була сильно підсилена. Якщо ми просто видалимо мінорну спеціальність то після нормування отримаємо:

$W^{***} = \left(\frac{0.471}{4410}, \frac{0.294}{4611}, \frac{0.235}{3508} \right)$. У цьому випадку не було ніякого підсилення для

спеціальності 3508 *Tourism*.

4.3 Контрольний приклад категоризації ключових слів

Розглянемо контрольний приклад категоризації ключових слів взятого з профілю потенційного рецензента. В демонстраційних цілях використаємо профіль потенційного рецензента з системи Google Scholar, в якій кожний науковець може вказувати свої наукові інтереси довільним чином, тобто у формі ключових слів. Тестовий приклад такого науковця подано на рис. 4.1. Для цієї дослідниці $w_1 = \text{"Information Retrieval"}$ та $w_2 = \text{"Human – Computer Interaction"}$. Порядок ключових слів у W немає значення, що відповідає моделі мішка слів. Згідно з алгоритмом додатково для цієї дослідниці згенеруємо третє ключове слово на основі поєднання двох існуючих: $w_3 = \text{"Information Retrieval" AND "Text Mining"}$. Якщо профіль містить 3 ключових слова то 3 додаткових ключових слова генеруються, якщо 4 тоді 6 додаткових тощо.



Susan Dumais

Technical Fellow, Microsoft Research
 Verified email at microsoft.com - [Homepage](#)
[Information Retrieval](#) [Human-Computer Interaction](#)

Рисунок 4.1 – Профіль тестового рецензента у Google Scholar

На рис. 4.2 подано початковий розподіл ступенів належності ключових слів до спеціальностей для профілю з рис. 4.1 (показано лише спеціальності, що мають ненульове значення по всім ключовим словам). Для кожного ключового слова а також попарного поєднання ключових слів ми шукаємо розподіл по спеціальностям науки з Dimensions. Після цього ми відкидаємо всі спеціальності з незначним ступенем належності встановленим на рівні *Tail* (рис. 4.3), що потім буде підбиратись в автоматичному режимі. Після цього усереднюємо розподіли ключових слів утворюючи лише один (рис. 4.4), і далі обмежуємо кількість ненульових належностей на рівні RG_max . Ми вважаємо, що рецензенти не можуть одночасно бути експертами у багатьох спеціальностях, тому також обмежуємо кількість спеціальностей до яких може бути категоризовано ключові слова на рівні не більше 4. Окрім цього також є обмеження по мінімальному рівні належності до спеціальності встановленому на рівні RG_min_degree . Це все подано на рис. 4.4 разом з результатом категоризації після усереднення, обмеження кількості ненульових спеціальностей на рівні RG_max і відкиданням спеціальностей з низьким рівнем належності, та кінцевий результат категоризації. Параметри *Tail*, RG_max , RG_min_degree будуть експериментально визначені в наступній секції. За результатами категоризації, тестову рецензентку на основі ключових слів з рис. 4.1 категоризовано до наступних спеціальностей:

4608 *Human-Centred Computing* зі ступенем належності 0.35;

4609 *Information Systems* зі ступенем належності 0.25;

4602 *Artificial Intelligence* зі ступенем належності 0.21;

4605 *Data Management and Data Science* зі ступенем належності 0.19.

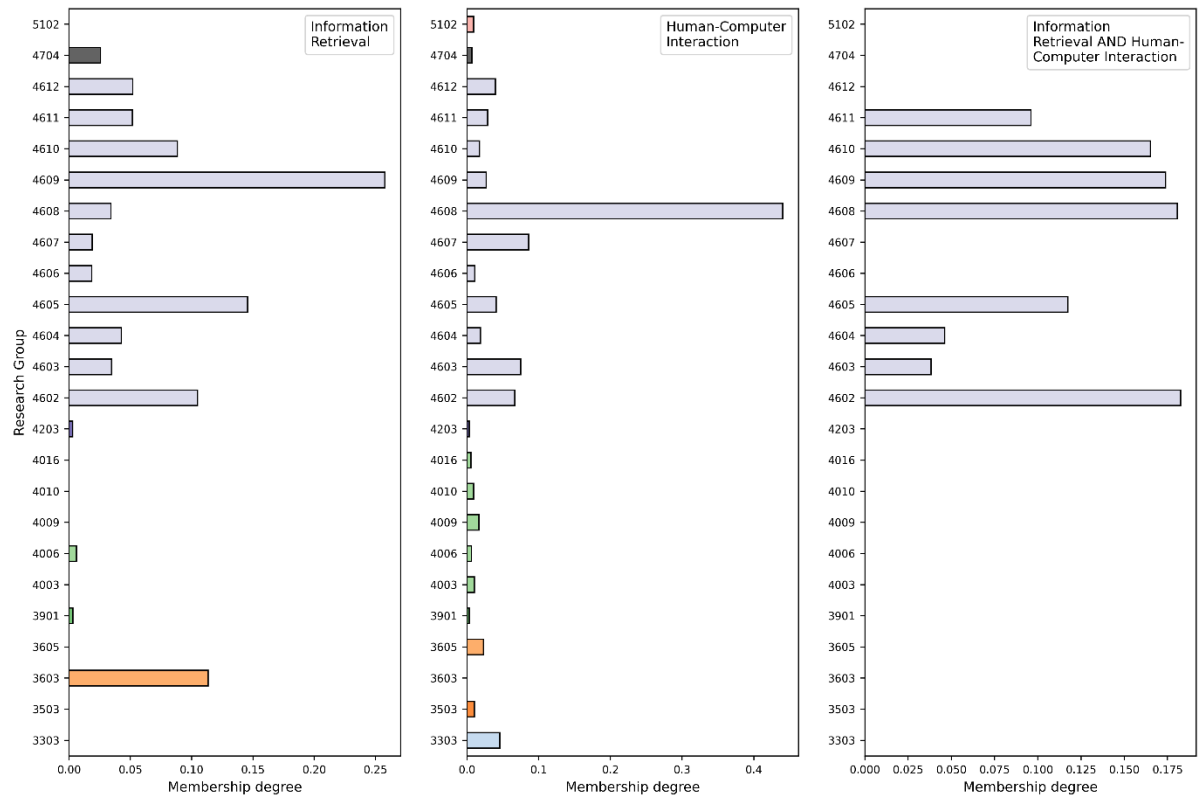


Рисунок 4.2 – Початковий розподіл ключових слів по науковим спеціальностям

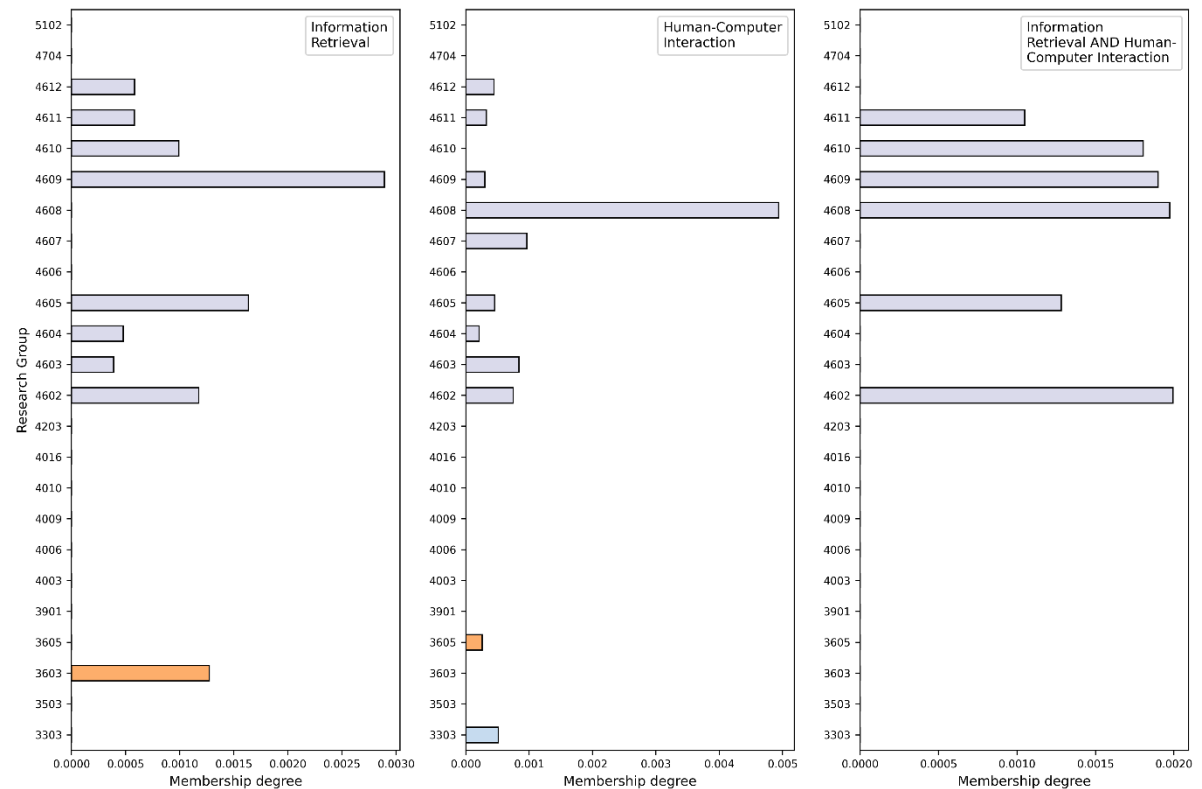


Рисунок 4.3 – Розподіл по спеціальностям після фільтрації

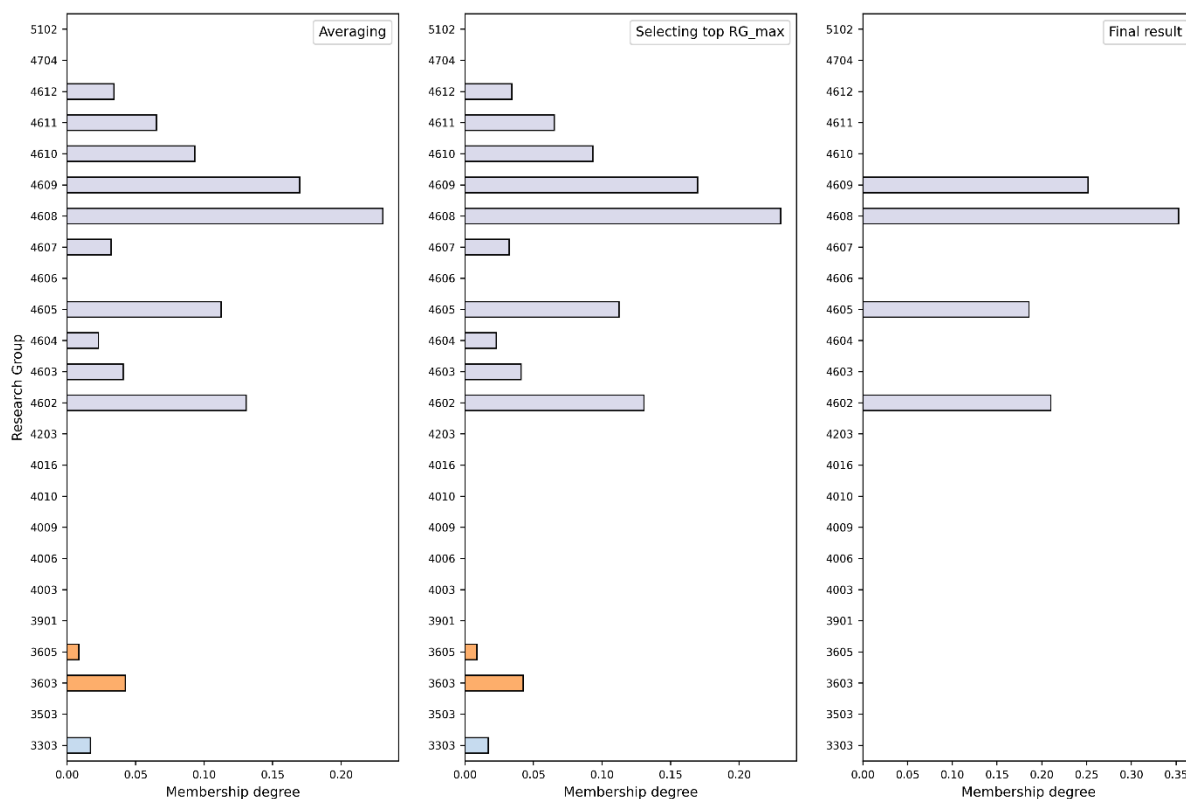


Рисунок 4.4 – Кінцеві розподіли тестової рецензентки

4.4 Підбір параметрів алгоритму категоризації

4.4.1 Ідея підбору параметрів алгоритму категоризації

Розглядається автоматичний підбір параметрів алгоритму категоризації ключових слів, а саме: *Tail*, *RG_max*, *RG_min_degree*. Їх необхідно обрати таким чином, щоб випадково не відкинути спеціальності, які відносяться до ключових слів. Ми пропонуємо це зробити на основі статей опублікованих потенційними рецензентами, а також потенційними авторами заявок на рецензування (в даному контексті це не важливо, так як алгоритм оперує суто ключовими словами). Ідея полягає в тому, щоб використати інформаційні ресурси системи Dimensions наступним чином. Використовуючи Dimensions можна знайти розподіл науковця за спеціальностям науки ANZSRC-2020 (рис. 4.5). Базуючись на цьому пропонується сформувати датасет потенційних рецензентів з набором ключових слів; ключові слова категоризувати відповідно до алгоритму запропонованого вище; категоризувати статті рецензентів використовуючи третій етап запропонованого алгоритму (*RG_max* та *RG_min_degree*

використовуються для цього); і зрештою порівняти категоризацію на основі ключових слів та на основі статей з Dimensions. В даному випадку ми припускаємо, що статті рецензента є джерелом про його сферу експертизи, і порівнюємо наскільки добре категоризація ключових слів співпадає з категоризацією за статтями.

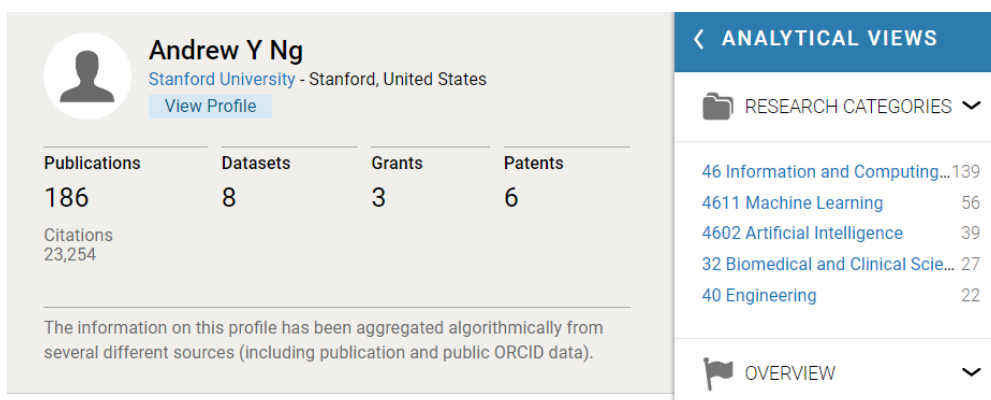


Рисунок 4.5 – Профіль науковця у Dimensions, з правого боку знаходиться розподіл по спеціальностям науки на основі категоризованих статей

4.4.2 Приклад використання запропонованого критерію схожості для порівняння розподілів на основі статей та на основі ключових слів

Для порівняння категоризації на основі ключових слів та на основі статей з Dimensions ми використовуємо критерій оцінювання схожості категоріальних розподілів запропонований у попередньому розділі. Розглянемо детальний приклад роботи критерію для тестової рецензентки з рис. 4.1. В результаті категоризації цієї рецензентки на основі її категоризованих публікацій в Dimensions отримуємо наступний розподіл:

4608 Human-Centred Computing з належністю 0.33,
4605 Data Management and Data Science з належністю 0.25,
4602 Information Systems з належністю 0.25,
4611 Machine Learning з належністю 0.17.

Для зручності представимо результати порівняння в табл. 4.1. З таблиці видно, що розподіл за інтересами та статтями має співпадіння у двох спеціальностях, а інші три спеціальності, які відрізняються, є спорідненими і

належать до однієї галузі науки з ANZSRC-2020. Це означає, що розподіл за інтересами певною мірою відображає тематику робіт її статей.

Таблиця 4.1 – Порівняння результатів категоризації для рецензентки Susan Dumais відповідно до запропонованого критерію оцінювання схожості

Спеціальність науки	Категоризація ключових слів з Google Scholar	Категоризація статей з Dimensions	Залишки від ключових слів	Залишки від статей
<i>4602 Artificial Intelligence</i>	0.21	0	0.21	0
<i>4605 Data Management and Data Science</i>	0.19	0.25	0	0.06
<i>4608 Human-Centred Computing</i>	0.35	0.33	0.02	0
<i>4609 Information Systems</i>	0.25	0.25	0	0
<i>4611 Machine Learning</i>	0	0.17	0	0.17

Відповідно до даних з табл. 4.1, отримуємо наступну оцінку схожості розподілів за критерієм (3.2): $F(Dumais) = 0.46$. Це перший доданок запропонованого критерію у формулі (3.1). Другий доданок передбачає врахування перехресної спорідненості між спеціальностями науки (формула (3.3)). Для цього необхідно порахувати залишки після визначення перетину двох розподілів, а також визначити спорідненість усіх пар спеціальностей до яких була категоризована рецензентка. Залишки подані у табл. 4.1, а спорідненість визначається використовуючи індекс Жаккара як показано в попередньому розділі. Для підрахунку індексу Жаккара дані з Dimensions беруться за 2019-2023 рр.. Для фільтрації шумів підрахунок перехресної спорідненості за формулою (3.3) здійснюємо лише для пар спеціальностей, які мають значний рівень схожості, з індексом Жаккара понад 0.05. Для спеціальностей з табл. 4.1 маємо лише 3 таких пари:

$$J(\langle 4602 \text{ Artificial Intelligence} \rangle, \langle 4611 \text{ Machine Learning} \rangle) = 0.09;$$

$$J\left(\langle 4602 \text{ Artificial Intelligence} \rangle, \langle 4605 \text{ Data Management and Data Science} \rangle\right) = 0.122;$$

$$J(\langle 4605 \text{ Data Management and Data Science} \rangle, \langle 4611 \text{ Machine Learning} \rangle) = 0.06.$$

Підрахунок залишків має наступний вигляд:

$$e_S(4602) = 0.21; \quad e_D(4602) = 0;$$

$$e_S(4611) = 0; \quad e_D(4611) = 0.17;$$

$$e_S(4605) = 0; \quad e_D(4605) = 0.06;$$

$$\varepsilon_S(4608) = 0.02; \quad \varepsilon_D(4608) = 0.$$

Підставляючи це все у формулу (3.3), отримуємо внески від перехресної спорідненості:

$$Fit(Dumais) = [0.09 \cdot \min(0.21, 0.17) + 0.09 \cdot \min(0, 0)] +$$

$$[0.06 \cdot \min(0, 0.17) + 0.06 \cdot \min(0, 0.06)] +$$

$$[0.122 \cdot \min(0.21, 0.06) + 0.122 \cdot \min(0, 0)]$$

$$Fit(Dumais) = 0.015 + 0 + 0.007 = 0.022$$

Беручи до уваги перехресну спорідненість, критерій схожості з формули (3.1) для тестового прикладу приймає наступне значення:

$$Fit(Dumais) = 0.77 + 0.02 = 0.79.$$

4.4.3 Формування датасету

Для підбору параметрів *Tail*, *RG_min_degree*, *RG_max* сформуємо датасет науковців. Для цього використаємо API Dimension і відсортуємо профілі науковців кількістю публікацій категоризованих до спеціальностей 4602 *Artificial Intelligence*, 4609 *Information Systems* та 4605 *Data Management and Data Science*. З 20000 науковців з найбільшою кількістю публікацій за 2018-2023 рр. оберемо тих, що задовольняють такі вимоги:

- 1) мають від 5 до 70 публікацій в Dimensions за 2018-2022 рр.;
- 2) мають щонайменше 2 інтереси в Google Scholar англійською мовою;
- 3) розподіли інтересів яких в Dimensions за спеціальностям науки не пусті;
- 4) є лише один профіль в Google Scholar за ім'ям з Dimensions.

Датасет очищено від авторів з статтями, розподіл яких розповсюджується на понад 15 спеціальностей в Dimensions. Також вилучено авторів, які категоризовані в Dimensions за абсолютно неспорідненими спеціальностями з

Функцій помилки враховує перехресну спорідненість для сильно схожих спеціальностей з індексом Жаккара на рівні 0.05. Таких пар спеціальностей всього є 134, включаючи ті до яких віднесені науковці з датасету:

4602 *Artificial Intelligence* та 4605 *Data Management and Data Science* зі схожістю 0.122;

4602 *Artificial Intelligence* та 4611 *Machine Learning* зі схожістю 0.09;

4609 *Information Systems* та 3503 *Business Systems in Context* зі схожістю 0.106.

Оптимізація виконана тричі з трьох різними цільовими функціями. Як кінцевий результат обирається середнє значення параметрів, отриманих в результаті кожної оптимізації. Використовуються наступні цільові функції:

$$\text{Mean}(A) = \frac{1}{|A|} \sum_{\forall a \in A} L(W_S^*(a), W_D^*(a)); \quad (4.3)$$

$$\text{Median}(A) = \text{median}_{\forall a \in A} (L(W_S^*(a), W_D^*(a))); \quad (4.4)$$

$$\text{Threshold}(A) = \frac{1}{|A|} \sum_{\forall a \in A} g(a). \quad (4.5)$$

де A – множина науковців разом з розподілами по інтересам та статтям;

$$g(a) = \begin{cases} 1, & \text{if } L(W_S^*(a), W_D^*(a)) < 0.5 \\ 0, & \text{otherwise} \end{cases} \quad \text{позначає, що категоризації для науковця } a$$

співпадають.

Параметри алгоритму підбираються у наступних межах:

[75, 98] для *Tail*;

[10, 20] для *RG_min_degree*;

[6, 15] для *RG_max*.

Підбір параметрів здійснюється на наступній сітці: для *Tail* – 10 таких дискретних значень {75, 80, 85, 88, 90, 94, 95, 96, 97, 98}; для *RG_min_degree* – 11 рівномірно розподілених значень; для *RG_max* – 10 рівномірно розподілених значень. Загальна кількість комбінацій параметрів – 1100. Складність цієї задачі оптимізації низька, тому можна використовувати простий пошук на сітці.

Результати пошуку на сітці подана на рис. 4.7. По вісі X подано ранговану множину можливих комбінацій параметрів. З рисунка видно, що помилки не зменшуються швидко для трьох цільових функцій. Найкращі параметри для цільових функцій (4.3) – (4.5) подані в табл. 4.2. Значення цільових функцій за оптимальних параметрів для кожного науковця подано на рис. 4.8. Більшість науковців мають помірний рівень співпадіння між розподілами на основі інтересів та розподілами статей, приблизно 2200 (89%) з них мають помилку менши за 0.8 взаємності від цільової функції. Кількість науковців з не нульовою схожістю між відповідними розподілами становить 2390 у випадку *Median* цільової функції, 2392 у випадку *Mean*, 2282 у випадку *Threshold* та 2370 (95%) для оптимальних параметрів. Кількість науковців при помилці $L(W_S^*, W_D^*) \leq 0.5$ становить 1479 для *Median*, 1473 для *Mean*, 1949 для *Threshold* та 1291 (52%) для оптимальних параметрів. Кількість науковців при помилці $L(W_S^*, W_D^*) \leq 0.33$ становить 702 для *Median*, 712 для *Mean*, 747 для *Threshold* та 692 (28%) для оптимальних параметрів.

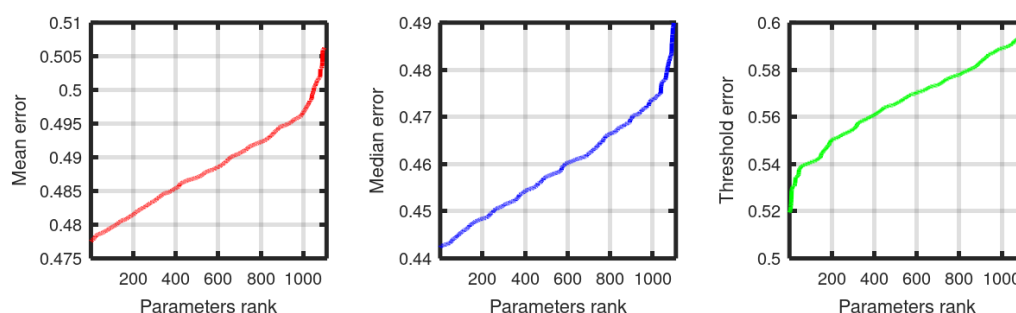


Рисунок 4.7 – Крива навчання параметрів алгоритму

Таблиця 4.2 – Оптимальні параметри для кожної цільової функції

Цільова функція	Tail	RG_min_degree	RG_max	Mean error	Median error	Threshold error
Mean Error	96	10	15	0.4775	0.4429	0.5926
Median Error	94	10	12	0.4781	0.4423	0.5950
Threshold Error	75	20	14	0.5060	0.4899	0.5195
Обрані	88	13	14	0.4846	0.4517	0.5761

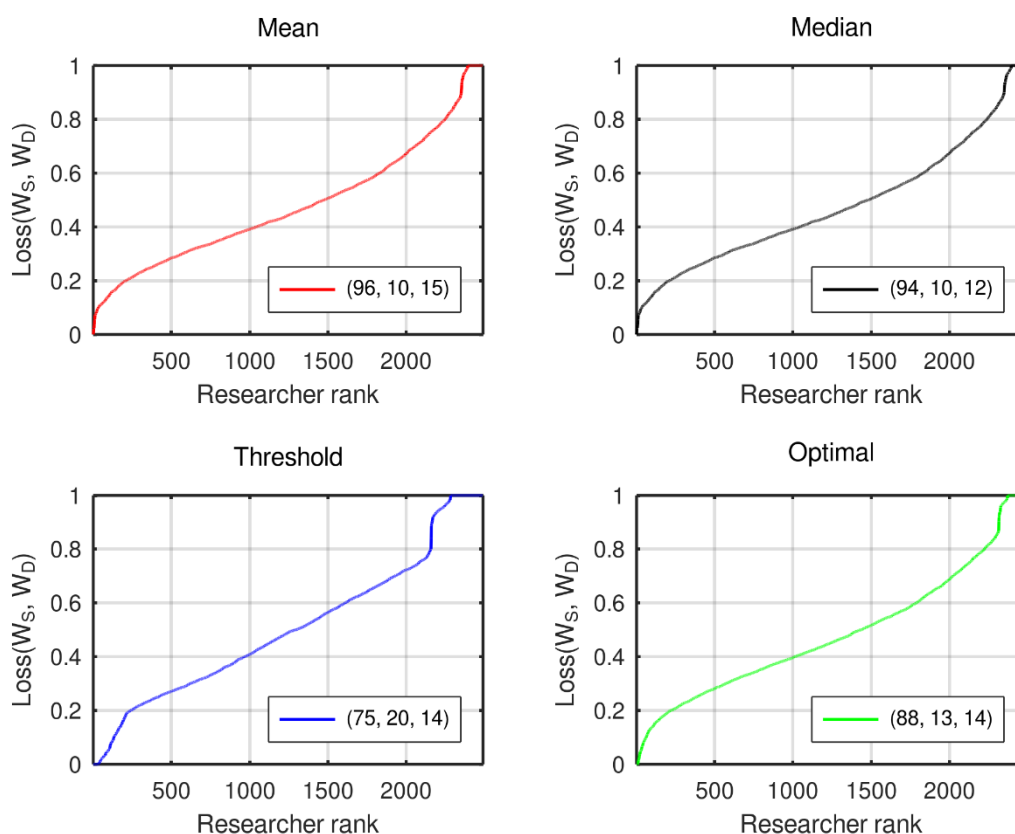


Рисунок 4.8 – Розподіл помилок для трьох цільових функцій; легенда містить оптимальні параметри у наступному форматі (*Tail*, *RG_min_degree*, *RG_max*)

В цілому розподіл помилок з оптимальними параметрами для трьох цільових функцій приблизно однаковий, але параметри трохи різняться. Це означає, що вибір цільової функції не має сильного впливу на результати оптимізації.

Отримані результати порівняння категоризації на основі інтересів та статей показують непоганий рівень співпадіння враховуючи те, що для такого типу задачі рідко можливо знайти відповідність між інтересами науковця та його статтями – список інтересів значно менший за кількість інформації у статтях науковця. Список інтересів – це зазвичай короткий опис специфічного галузевого знання, яке є низькорівневим, в той час як науковці можуть писати статті які вирішують проблеми з високорівневими концепціями і використовують низькорівневі поняття. Тому статті науковця можуть бути категоризовані не до спеціальностей які представляють його/її низькорівневі інтереси, а до спеціальностей які пов'язані з високорівневими поняттями

проблем, що вирішуються у статтях науковця. Деякі статті можуть мати міждисциплінарну природу і зазвичай пишуться кількома науковцями; кожний науковець може робити внесок лише у межах спеціальності в якій він експерт, що означає що для міждисциплінарних статей важко визначити внесок кожного науковця окремо.

Для прикладу, результати порівняння чотирьох науковців з найкращим, посереднім та найгіршим співпадінням розподілів категоризації на основі інтересів та статей подано у табл. 4.3. Повні назви спеціальностей з табл. 4.3 наведено у Додатку Б.

Таблиця 4.3 – Фрагмент порівняння результатів категоризації науковців з оптимальними параметрами

Ранг	Ім'я	Інтереси (без попарного поєднання)	Категоризація на основі інтересів	Категоризація на основі статей	Помилка
1	Kening Zhu	Tangibles, Wearables, Mobiles, Rapid Prototyping, Human Computer Interaction.	4608 -1.0	4608 - 1.0	0
2	Eduard Gröller	Visualization, Computer Graphics, Visual Computing.	4607 – 1.0	4607 – 1.0	0
3	Danilo Gallo	Human-Robot Interaction, User Experience Design, Human-Computer Interaction.	4608 – 1.0	4608 – 1.0	0
4	Yeliz Yesilada	Human Computer Interaction, Web accessibility, Accessibility, World Wide Web, Eye-Tracking.	4608 – 1.0	4608 – 1.0	0
1245	Dominic Rosch-Grace	Data Driven Control, System Identification, Machine Learning, Dynamic Networks, Estimation.	4611 – 0.3290 4605 – 0.3012 4007 – 0.2066 4010 – 0.1631	4007 – 0.4527 4010 – 0.2382 4611 – 0.1786 4901 – 0.1305	0.4516
1246	Juan-Manuel Torres-Moreno	Artificial Consciousness, Language Engineering, Text Summarization, Text Classification.	4602 – 0.4155 4608 – 0.2010 4605 – 0.1935 4612 – 0.1809	4605 – 0.4665 4602 – 0.3457 4704 – 0.1878	0.4521

1247	De-Nian Yang	Network Analysis and Optimization, Social and Multimedia Networks.	4605 – 0.3273 4701 – 0.2653 4603 – 0.2183 4611 – 0.1891	4605 – 0.3461 4606 – 0.2615 4603 – 0.2290 4608 – 0.1634	0.4524
1248	Piljoo Choi	Cryptography, Hardware design.	4604 – 0.6704 4606 – 0.1815 4009 – 0.1481	4604 – 0.3789 4009 – 0.3096 4605 – 0.1611 4006 – 0.1502	0.4528
2486	Monique van Velzen	Pain, patient phenotyping.	3204 – 0.2970 3205 – 0.2722 3209 – 0.2241 4201 – 0.2067	3214 – 0.5241 3202 – 0.4759	1
2487	Wissam Al Jurdi	Artificial Intelligence, Machine Learning, Data Science, Recommender Systems.	4602 – 0.3246 4608 – 0.2470 4611 – 0.2264 4605 – 0.2019	4609 – 1.0	1
2488	Oliver Bimber	Computer Graphics, Computer Vision, Visual Computing, Visualization, Displays.	4607 – 0.8083 4608 – 0.1917	4013 – 0.6214 4605 – 0.2450 4603 – 0.1336	1
2489	Pierre-Cyrille Heam	Finite Automata, Software Verification and Testing.	4612 – 0.7202 4613 – 0.2798	4606 – 0.4882 4605 – 0.1721 4602 – 0.1770 4601 – 0.1627	1

Науковці у табл. 4.3 відсортовані зі збільшенням помилки, значна частина з них категоризована до галузей *40 Engineering* та *46 Information and Computing Sciences*. Перших 4 науковців мають однакові результати категоризації на основі інтересів та статей. Середні науковці мають значення помилки приблизно на рівні 0.45, серед них є співпадіння категоризації по спеціальностях та галузях. Науковці з нульовим співпадінням категоризацій мають спеціальності інтересів та статей з однакових або різних галузей науки. Розглянемо деякі з них. У Monique van Velzen інтереси пов'язані з *pain* та *patient phenotyping*, в той час як її статті сфокусовані на створенні та аналізі ліків, тому розподіли різні, але більшість спеціальностей належать до однієї галузі знань. Oliver Bimber має інтереси пов'язані з комп'ютерною графікою, а статті стосуються застосування комп'ютерної графіки для аналізу поверхні землі використовуючи зображення з дронів, тому розподіли різні – дві спеціальності належать до однієї галузі і пов'язані з комп'ютерною графікою,

одна спеціальність пов'язана з *4013 Geomatic Engineering*. Інтереси Wissam Al Jurdi є близькими до *4602 Artificial Intelligence*, *4611 Machine Learning*, *4605 Data Management and Data Science*, *4608 Human-Centred Computing*, в той час як його статті пов'язані з *4609 Information Systems*, тому що зміст статей стосується в основному рекомендаційних систем, обидві групи спеціальностей пов'язані і належать до однієї галузі науки. Ці приклади показують, що в більшості випадків інтереси науковця відповідають статтям які він/вона написали (близько 89% з них мають помилку менше за 0.8), і також є рідкі випадки коли інтереси не достатньо добре співставляються з вмістом статей.

4.5 Категоризація за мішком ключових слів

Розглянемо випадок, коли потенційний рецензент описується не однією множиною інтересів, а кількома. В теорії класифікації це відповідає навчанню за набором зразків (multiple instance learning) [171]. Наприклад, рецензент є автором кількох статей, кожна з яких має свій набір ключових слів. І за цими статтями потрібно категоризувати рецензента до однієї чи кількох спеціальностей.

В підр. 4.1 розглянута ситуація категоризації ключових слів статей під час формування датасету для налаштування параметрів запропонованого алгоритму категоризації. Тоді, під час формування датасету використовувалися результати категоризації статей, яка здійснена системою Dimensions. Для опрацювання категоризації статей з Dimensions використовувався третій етап запропонованого алгоритму. Нижче пропонується алгоритм такої категоризації, коли відомі лише ключові слова кожної статті.

Для категоризації декількох статей необхідно:

- 1) для кожної статті сформувати множину ключових слів з попарним поєднанням ключових слів;
- 2) об'єднати отримані множини ключових слів в одну множину;

3) виконати категоризацію отриманої множини ключових слів за запропонованим алгоритмом категоризації ключових слів.

4.6 Підбір колективу рецензентів

4.6.1 Загальна постановка задачі підбору рецензентів

Розглядається задача підбору колективу рецензентів, які сукупно найкраще підходять для експертизи заявки. При цьому можливі 2 випадки: формування колективу з нуля та доповнення до колективу. Прикладом такої задачі є формування разової спецради із захисту PhD-дисертації.

Вважатимемо відомими:

$P = \{p_1, p_2, \dots, p_n\}$ – профіль заявки, що визначений множиною її ключових слів;

$P^* = \{\mu_P(t_1), \mu_P(t_2), \dots, \mu_P(t_m)\}$ – профіль заявки у формі категоріального розподілу за спеціальностям, де $\mu_P(t_i) \in [0, 1]$ – ступінь належності профілю P до спеціальності t_i , $i = \overline{1, m}$;

$R = \{r_1, r_2, \dots, r_k\}$ – множина профілів потенційних рецензентів, з яких необхідно обрати тих хто рецензуватиме P , де $r_i = \{r_{w1}, r_{w2}, \dots, r_{wl}\}$ – профіль рецензента у вигляді сукупності множин ключових слів, $i = \overline{1, k}$;

$R^* = \{\mu_1, \mu_2, \dots, \mu_k\}$ – множина профілів потенційних рецензентів в просторі спеціальностей;

$\mu_i = \{\mu_R(t_1), \mu_R(t_2), \dots, \mu_R(t_m)\}$ – ступінь належності профілю рецензента r_i до спеціальності t_j , $i = \overline{1, k}$, $j = \overline{1, m}$.

Необхідно знайти:

$$S \in R^* = \left\{ P^*, \{\mu_1, \dots, \mu_C\} \in R^* \mid \max \left(\text{Fit}(P^*, \text{Aggregate}(\{\mu_1, \dots, \mu_C\})) \right) \right\} -$$

підмножину рецензентів, що мають найбільшу сукупну відповідність заявці, де $\text{Fit}(P^*, \text{Aggregate}(\{\mu_1, \dots, \mu_C\}))$ – рівень відповідності підмножини потенційних

рецензентів спеціальностям заявки, який розраховано за запропонованим критерієм схожості категоріальних (3.1); $Aggregate(\{\mu_1, \dots, \mu_C\})$ – функція агрегації категоріальних розподілів підмножини потенційних рецензентів, $2 \leq C \leq 5$ – кількість рецензентів необхідних для рецензування.

За такої постановки задачі завантаженість рецензентів не враховується, тобто рецензент може бути призначений одночасно до декількох заявок.

Частковим випадком задачі підбору рецензентів є задача підбору опонентів разової ради до раніше обраних рецензентів та голові разової ради. Ця задача є випадком коли $C \in \{2, 3\}$ в залежності від того, скільки потрібно опонентів підібрати для дисертації. Для автоматичного підбору опонентів можуть використовуватись ключові слова дисертації та ключові слова статей потенційних опонентів.

4.6.2 Алгоритми вирішення задачі підбору рецензентів

Задача підбору рецензентів є NP-повною. Для таких задач, алгоритми вирішення які застосовуються на практиці, є переважно наближеними. Нижче подано деякі алгоритми її вирішення. В кінцевому рахунку необхідно обрати той, який забезпечує баланс за критеріями якості підбору – витрати ресурсів.

4.6.2.1 Повний перебір

Оптимальний варіант можна знайти повним перебором: для заявки P^* необхідно серед усіх можливих комбінацій множини потенційних рецензентів $\{\mu_1, \dots, \mu_C\}$ обрати таку, що забезпечує максимальне значення відповідності $Fit(P^*, Aggregate(\{\mu_1, \dots, \mu_C\}))$. Складність такого перебору зростає експоненційно, тому для великих розмірів даних перебрати всі можливі варіанти та вкласти в якісь обмеження по часу є нереальним.

4.6.2.2 Повний перебір на обрізаній множині кандидатів

Повний перебір враховує всі можливі комбінації потенційних кандидатів розміру C . Хоча на практиці кандидати з низьким рівнем схожості до заявки наврядчи будуть призначені до неї. В зв'язку з цим виникає питання визначення балансу «рівень відповідності – тривалість перебору». Тобто як

зміниться рівень відповідності сукупності підібраних рецензентів якщо знехтувати деякими потенційними рецензентами з низькою схожістю і таким чином скоротити. Для цього можна виконувати повний перебір, але за обрізаним (truncated) списком кандидатів, наприклад, відкинути потенційних рецензентів з низьким рівнем схожості, скажімо, 0.1, 0.2 чи 0.3.

4.6.2.3 Жадібний алгоритм

Існують наближені алгоритми підбору рецензентів, що не гарантують оптимального розв'язку але при цьому можуть значно пришвидшити сам процес піжбору. Один з варіантів – це жадібний алгоритм. Його суть полягає у тому, що рецензенти підбираються ітеративно з забезпечення максимальної відповідності поточного складу колективу рецензентів. Алгоритм складається з S ітерацій. На першій ітерації знаходиться кандидат, що має найвищу схожість з дисертацією. На другій ітерації обираємо кандидата, що сукупно з уже підібраним на першій ітерації кандидатом має найвищу відповідність з дисертацією і т.д. Кількість операцій при такому підході значно зменшується, і відповідно скорочується тривалість перебору. При цьому розв'язок може вийти неоптимальним. У випадку задачі доповнення колективу рецензентів можливі 2 варіанти жадібного алгоритму: перший виконує максимізацію відповідності колективу утвореного з додаванням необхідної кількості рецензентів, а другий спочатку додає найбільш схожого кандидата і далі максимізує відповідність колективу додаючи найбільш підходящого кандидата. Другий варіант є реалізацією концепції елітизму з метаевристичних алгоритмів оптимізації.

4.6.2.4 Ізольований підбір

Найпростіший спосіб призначити рецензентів це обрати тих, що найбільш схожі до профілю дисертації. Цей підхід буде лінійним по швидкості, але не гарантуватиме оптимального розв'язку так як наявність найбільш схожих рецензентів не означає, що усі спеціальності дисертації будуть повністю покриті спеціальностями рецензентів. Суть полягає в тому, що кандидатів сортують по спаданню рівня схожості до дисертації та обирають рецензентами перших S кандидатів.

4.6.3 Оцінювання відповідності колективу рецензентів до заявки

Розглянемо приклад визначення відповідності разової ради до дисертації. Для цього використаємо інформаційну систему НАЗЯВО [172], у якій зберігаються дані за захистами PhD-дисертацій. Для прикладу фрагмент даних тестової дисертації з сайту НАЗЯВО подано на рис. 4.9.

Відомості про захист

ID 4136

Ступінь присуджено

Подати повідомлення

Загальні відомості

ПІБ здобувача	Пуріш Сергій Володимирович
Заклад освіти, освітня програма	Національний університет "Одеська політехніка", ОП 50837 Комп'ютерні науки (122 Комп'ютерні науки)
Відомості про акредитацію освітньої програми (за даними ЄДЕБО)	Акредитація освітньої програми (Національне агентство), сертифікат № 4654, діє до 01.07.2027
Освітня програма реалізовується у співпраці з іншим закладом освіти	ні

[Відомості про ОП \(NAQA.Registry\)](#)

Дисертація

Тема дисертації	Методи машинного навчання для розпізнавання людини за ходом.
-----------------	--

Рисунок 4.9 – Тестова дисертація з інформаційної системи НАЗЯВО

Автор дисертації – *Пуріш Сергій Володимирович*, назва – *Методи машинного навчання для розпізнавання людини за ходом*. Для оцінювання відповідності спецради дисертації використовуємо ключові слова дисертації англійською мовою. Для тестової дисертації вони мають наступний вигляд:

$$P = \left\{ \begin{array}{l} intelligent\ systems, machine\ learning, biometric\ identification, \\ neural\ networks, computer\ vision, recognition, classification, \\ identification, digital\ image\ processing, texture, contour, \\ descriptors, recognition\ efficiency, iterative\ algorithms \end{array} \right\}.$$

Виконаємо категоризацію поданих ключових слів використовуючи запропонований алгоритм категоризації. В результаті отримаємо:

$$P^* = \left(\frac{0.628}{4603}, \frac{0.207}{4611}, \frac{0.164}{4605} \right). \text{ Ключові слова було категоризовано до таких}$$

спеціальностей: 4611 *Machine Learning*, 4603 *Computer Vision and Multimedia Computation*, 4605 *Data Management and Data Science* . Дисертація була захищена по спеціальності 122 *Комп'ютерні науки*.

Аналогічним чином категоризуємо профілі членів разової ради використовуючи ключові слова з їх статей, які також зазначені в інформаційній системі НАЗЯВО. Вони є наступними.

Голова: а) *Computer vision, neural networks, dataset, transformer, action understanding, video understanding, artificial intelligence based proctoring systems, online proctoring, proctoring system, distance learning, online learning*; б) *Biometric Face Recognition, Convolutional Neural Network, Deep Learning, Computer Vision, Face Detection, Haar Cascade, Image Processing, Face Dataset*; в) *Computer vision, neural networks, deep learning, metrics, video processing, 2D, efficiency*.

Рецензент 1: а) *pseudo-labeling data, semi-supervised learning, transfer learning, convolution neural networks, facial emotion recognition*; б) *Convolution neural networks, facial expression recognition, deep learning, transfer learning, multilabel classification*; в) *Deep Learning, Transfer Learning, Facial Expression Recognition, Convolutional Neural Networks*.

Рецензент 2: а) *nonlinear dynamic objects, diagnostic models, model reduction, feature selection, correlation analysis*; б) *statistical characteristics, random variable, power transformations, machine learning models*; в) *identification, nonlinear objects, significant nonlinearities, dynamic neural networks, simulation modeling*.

Опонент 1: а) *Technical diagnostics, non-destructive diagnostic methods, regeneration, digital image processing, the Zernike moments, defect, microwave engineering*; б) *COVID-19, machine learning, data preprocessing, feature selection, classification*; в) *Group Method of Data Handling, Iterative Algorithm, Gross Domestic Product, Investment, Model Adequacy*.

Опонент 2: а) *methods of recognition stereo pair, depth map, correlation algorithm, perturbation functions, operation of subtraction*; б) *virtual environment,*

plenoptic function, image-based rendering, visualization of complex scenes, computer vision; в) testing, Hesberg's method, realism, rasterization, perspective-correct overlaying of textures.

Ключові слова статей членів ради категоризуються за принципом мішка ключових слів описаним у підрозділі 4.5. Результати категоризації зведено у табл. 4.4. Для трьох членів ради рівень схожості нижче 0.5, для двох інших схожість не перевищує 0.75.

Повні назви спеціальностей з табл. 4.3 мають наступний вигляд:

3102 Bioinformatics and Computational Biology;

4005 Civil Engineering;

4007 Control Engineering, Mechatronics and Robotics;

4603 Computer Vision and Multimedia Computation;

4605 Data Management and Data Science;

4607 Graphics, Augmented Reality and Games;

4608 Human-Centred Computing;

4611 Machine Learning.

Таблиця 4.4 – Результати категоризації профілів членів разової ради

Член разової ради	Ступіні належності профілю члена до спеціальностей, μ_i	Схожість між дисертацією та членом ради $Fit(P^*, \mu_i)$
Голова	$\left(\frac{0.526}{4603}, \frac{0.276}{4611}, \frac{0.198}{4608} \right)$	0.743
Рецензент 1	$\left(\frac{0.430}{4611}, \frac{0.423}{4608}, \frac{0.147}{4603} \right)$	0.383
Рецензент 2	$\left(\frac{0.460}{4611}, \frac{0.216}{4605}, \frac{0.191}{3102}, \frac{0.132}{4007} \right)$	0.397
Опонент 1	$\left(\frac{0.298}{4611}, \frac{0.261}{4605}, \frac{0.232}{4005}, \frac{0.207}{4603} \right)$	0.592
Опонент 2	$\left(\frac{0.549}{4607}, \frac{0.184}{4605}, \frac{0.136}{4603}, \frac{0.132}{4608} \right)$	0.303

Для визначення відповідності тематики дисертації разовій раді необхідно агрегувати результати категоризації усіх членів ради. Тобто, необхідно визначити функцію агрегації розподілів членів разових рад:

$$Aggregate \left(\left\{ \left(\frac{0.526}{4603}, \frac{0.276}{4611}, \frac{0.198}{4608} \right), \left(\frac{0.430}{4611}, \frac{0.423}{4608}, \frac{0.147}{4603} \right), \right. \right. \\ \left. \left(\frac{0.460}{4611}, \frac{0.216}{4605}, \frac{0.191}{3102}, \frac{0.132}{4007} \right), \left(\frac{0.298}{4611}, \frac{0.261}{4605}, \frac{0.232}{4005}, \frac{0.207}{4603} \right), \right. \\ \left. \left(\frac{0.549}{4607}, \frac{0.184}{4605}, \frac{0.136}{4603}, \frac{0.132}{4608} \right) \right\} \right).$$

Дану функцію реалізуємо як третій етап алгоритму категоризації ключових слів. В результаті застосування третього етапу алгоритму категоризації для агрегування отримаємо наступні ступені належностей разової ради до спеціальностей: $\left(\frac{0.374}{4611}, \frac{0.263}{4603}, \frac{0.193}{4608}, \frac{0.170}{4605} \right)$. Відповідність разової ради дисертації за запропонованим критерієм (3.1) становить:

$$Fit \left(\left(\frac{0.628}{4603}, \frac{0.207}{4611}, \frac{0.164}{4605} \right), \left(\frac{0.374}{4611}, \frac{0.263}{4603}, \frac{0.193}{4608}, \frac{0.170}{4605} \right) \right) = 0.649.$$

Отримане значення можна змістовно інтрепретувати як рівень відповідності вище середнього.

Висновки до розділу 4

1. Розроблено метод категоризації рецензентів, який базується на перетворенні обмеженого набору початкової інформації у формі наукових інтересів рецензента або ключових слів заявки у вектор в просторі наукових спеціальностей за принципом word2vec, і відрізняється від формального співставлення шифрів спеціальностей аналізом фактичного змісту заявки та поточної компетентності рецензента, а від моделі співставлення ключових слів – врахуванням синонімії та спорідненості. Покроковий контрольний приклад продемонстрував, що результати категоризації відповідають експертним уявленням.

2. Для підбору параметрів алгоритму категоризації ключових слів сформовано тестовий датасет обсягом у 2489 науковців. Експерименти на тестовому датасеті показують, що близько 95% розподілів науковців за інтересами мають відмінну від нуля схожість із розподілами науковців за статтями. 95% є хорошим збігом через те, що інтереси науковця актуальні лише в певний момент часу, а його роботи можуть бути пов'язані з іншими спеціальностями науки. Кількість науковців з відмінною відповідністю становить 28%. Це досить високий показник для такого типу вхідної інформації, отже запропонований алгоритм категоризації може бути корисним для швидкого створення короткого списку кандидатів для проблем призначення рецензента та подібних завдань.

3. Запропоновано алгоритм категоризації рецензентів для випадку, коли список інтересів описується не однією множиною, а декількома, наприклад, коли для прийняття рішення беруться ключові слова кількох його статей.

4. Виконано загальну постановку задачі експрес-підбору колективу рецензентів. Запропоновано алгоритм категоризації колективу рецензентів. Запропоновано критерій відповідності колективу рецензентів до тематики заявки. Запропоновано 5 алгоритмів підбору колективу рецензентів, які відрізняються тривалістю перебору та якістю знайденого рішення.

Результати розділу опубліковано в роботах автора [12], [13], [14], [15], [17].

РОЗДІЛ 5

ПРОГРАМНА РЕАЛІЗАЦІЯ ТА ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ

5.1 Програмне забезпечення інформаційної технології

5.1.1 Архітектура системи

Розроблені в попередніх розділах критерії, моделі та алгоритми реалізовано програмно. Основні програмні компоненти написано на мові програмування PHP, аналіз даних з візуалізаціями виконано на мові програмування Python, підбір параметрів алгоритму категоризації ключових слів було здійснено на мові програмування Octave. Основні розроблені програмні компоненти (рис. 5.1) є такими:

- програмний компонент парсингу даних розподілів ключових слів по спеціальностям науки з Dimensions;
- програмний компонент парсингу розподілу кількості статей по спеціальностям;
- програмний компонент категоризації ключових слів до спеціальностей ANZSRC-2020;
- програмний компонент запропонованого критерію оцінювання схожості категоріальних розподілів;
- програмний компонент визначення спорідненості між двома спеціальностями з ANZSRC-2020;
- програмний компонент формування тестових даних для підбору параметрів алгоритму категоризації ключових слів;
- програмний компонент для оптимізації параметрів алгоритму категоризації ключових слів;
- програмний компонент аналізу датасету дисертацій;
- програмний компонент оптимізації призначень опонентів до дисертацій;

- програмний компонент візуалізації отриманих результатів.

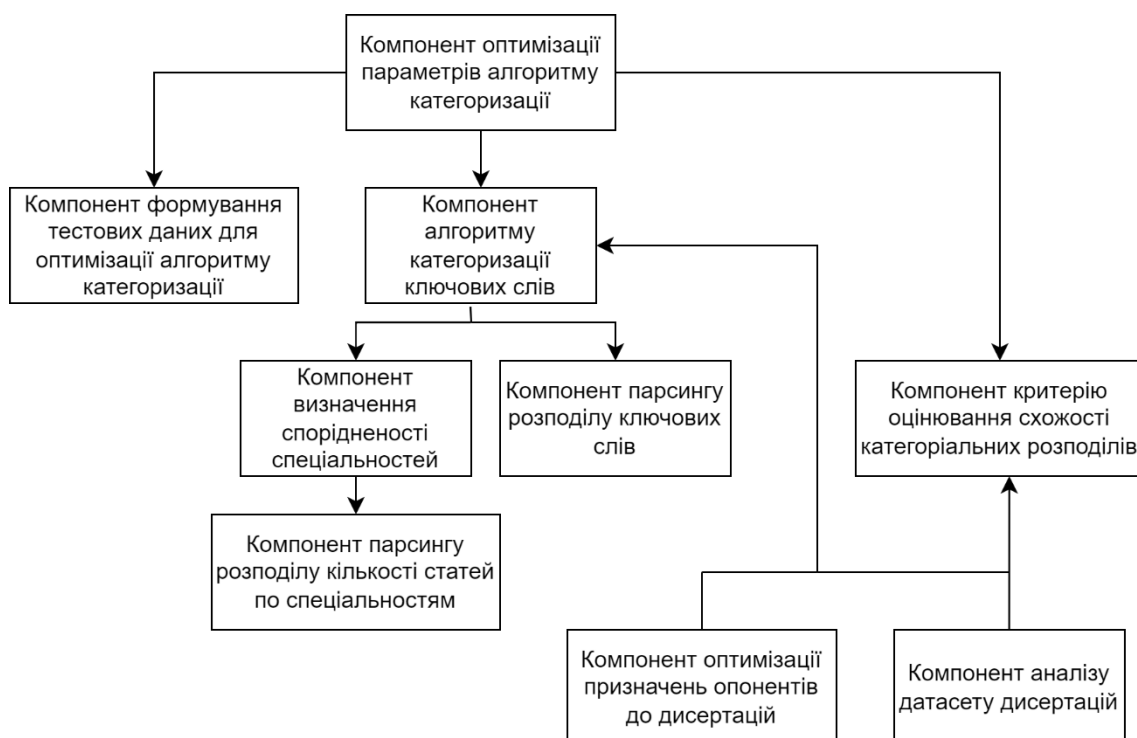


Рисунок 5.1 – Високорівнева архітектура взаємодії між розробленими програмними компонентами

5.1.2 Програмний компонент парсингу даних щодо розподілів ключових слів

Даний програмний компонент виконує парсинг даних розподілів ключових слів за допомогою системи Dimensions. Доступ до її ресурсів було надано *Digital Science & Research Solutions Inc.* в рамках проєкту *DIM-371*. Розподіл кількості публікації за спеціальностям науки, в яких згадується певне ключове, отримується через [158]. Для цього використовується звичайний HTTP запит, де тілом є текст у форматі Dimensions Search Language [159]. Наприклад, для того щоб отримати розподіл кількості публікацій за кожною спеціальністю, в яких зустрічається ключове слово «fuzzy logic» за останні 5 років необхідно відправити запит з таким тілом:

```
search publications
in title_abstract_only for "fuzzy logic"
```

```

where year in
[2019, 2020, 2021, 2022, 2023] return
category_for_2020

```

Для отримання загального розподілу кількості публікацій по всіх спеціальностям науки необхідно виконати наступний запит:

```

search publications
where year in
[2019, 2020, 2021, 2022, 2023] return
category_for_2020

```

Результат повертається у JSON-форматі, в якому для кожної спеціальності вказано кількість публікацій категоризованих до відповідної спеціальності. Для того, щоб знайти кількість публікацій віднесених одночасно до двох спеціальностей у запиті необхідно додатково вказати ідентифікатор спеціальності (для визначення схожості між спеціальностями). Як відповідь повертається кількість публікацій, які категоризовані одночасно до вказаної спеціальності і до кожної іншої спеціальності. Приклад такого запиту для спеціальності 3202, яка має внутрішній ідентифікатор 80045, наведено нижче:

```

search publications
where year in
[2019, 2020, 2021, 2022, 2023] and
category_for.id=80045
return category_for_2020

```

5.1.3 Програмний компонент категоризації ключових слів

Програмний компонент категоризації ключових слів є одним із головним в запропонованій інформаційній технології. На рис. 5.2 та 5.3 подані блок-схеми роботи 2 та 3 етапу алгоритму категоризації (перший етап доволі простий). Сам псевдокод подано в підр. 4.2. На блок-схемах наведена його високорівнева структура, а реалізація здійснена на мові програмування РНР. Другий етап для сукупності ключових слів (запитів) виконує пошук розподілу за спеціальностями та на виході для кожного ключового слова повертає

розподіл. На третьому етапі виконується агрегація розподілів за усіма ключовими і на виході отримуємо розподіл сукупності ключових слів за спеціальностям.

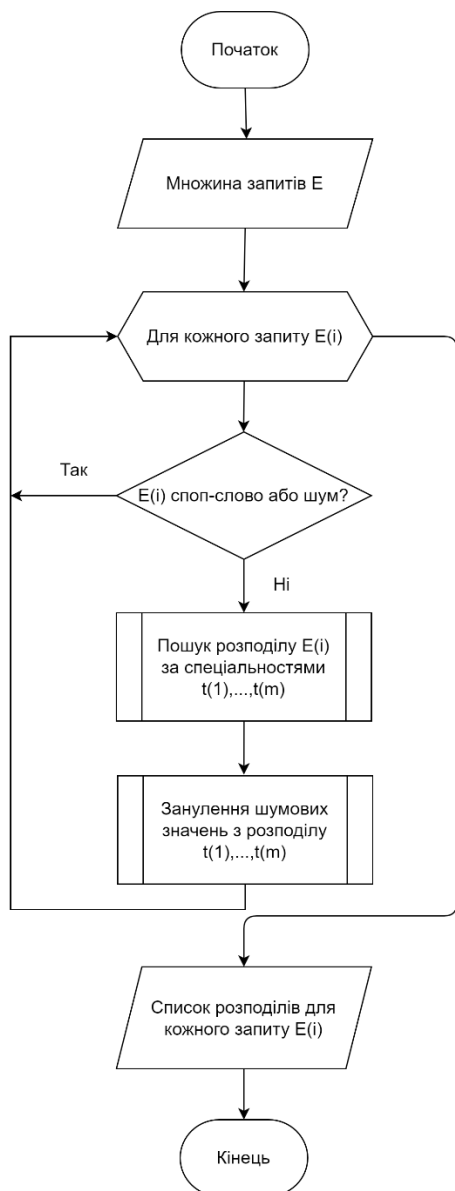


Рисунок 5.2 – Блок-схема другої ділянки алгоритму категоризації ключових слів



Рисунок 5.3 – Блок-схема третьої ділянки алгоритму категоризації ключових слів

UML-діаграма послідовності роботи компоненту категоризації ключових слів подана на рис. 5.4. На ній показано взаємодію з іншими програмними компонентами. Для отримання розподілів ключових слів за спеціальностями

використовується компонент парсингу розподілу ключових слів, а для визначення схожості між двома спеціальностями використовується компонент визначення схожості спеціальностей.

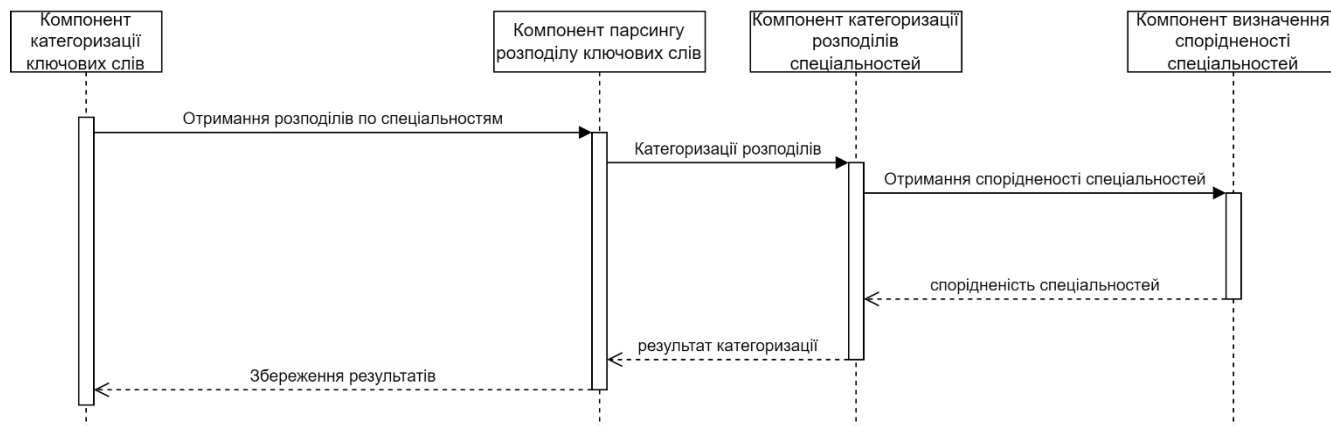


Рисунок 5.4 – UML-діаграма послідовності роботи програмного компоненту категоризації ключових слів

5.1.4 Програмний компонент критерію оцінювання схожості категоріальних розподілів

Програмний компонент критерію оцінювання схожості категоріальних розподілів є доволі простим. Для його роботи потрібно лише мати матрицю схожості спеціальностей, що знаходиться за допомогою компонента визначення схожості двох спеціальностей. UML-діаграма послідовності роботи цього компоненту подана на рис. 5.5.

5.1.5 Програмний компонент формування тестових даних для оптимізації параметрів алгоритму категоризації ключових слів

Для підбору параметрів алгоритму категоризації ключових слів сформовано датасет в автоматичному режимі. Початковий список науковців для датасету отримано з Dimensions, а ключові слова відповідних науковців отримано з Google Scholar. З використанням API Dimensions профілі науковців, яких категоризовано до спеціальностей 4602 Artificial Intelligence, 4609 Information Systems та 4605 Data Management and Data Science, відсортовано за

кількістю публікацій. UML-діаграма послідовності роботи цього компоненту подана на рис. 5.6.

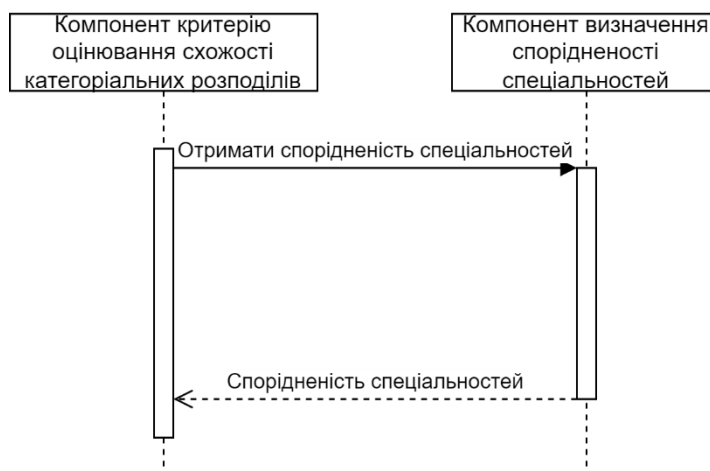


Рисунок 5.5 – UML-діаграма послідовності роботи програмного компоненту критерію оцінювання схожості категоріальних розподілів

5.1.6 Програмний компонент оптимізації параметрів алгоритму категоризації ключових слів

Для оптимізації параметрів алгоритму категоризації ключових слів розроблено програмний компонент, який працює з сформованим датасетом науковців. Для роботи цього компонента необхідний програмний компонент категоризації ключових слів та програмний компонент критерію оцінювання схожості категоріальних розподілів. Для оптимізації виконується категоризація ключових слів науковців з датасету та категоризація статей науковців, отримані результати порівнюються використовуючи критерій оцінювання схожості категоріальних розподілів та на основі цього шукається оптимальна комбінація параметрів. UML-діаграма послідовності роботи цього компоненту подана на рис. 5.7.

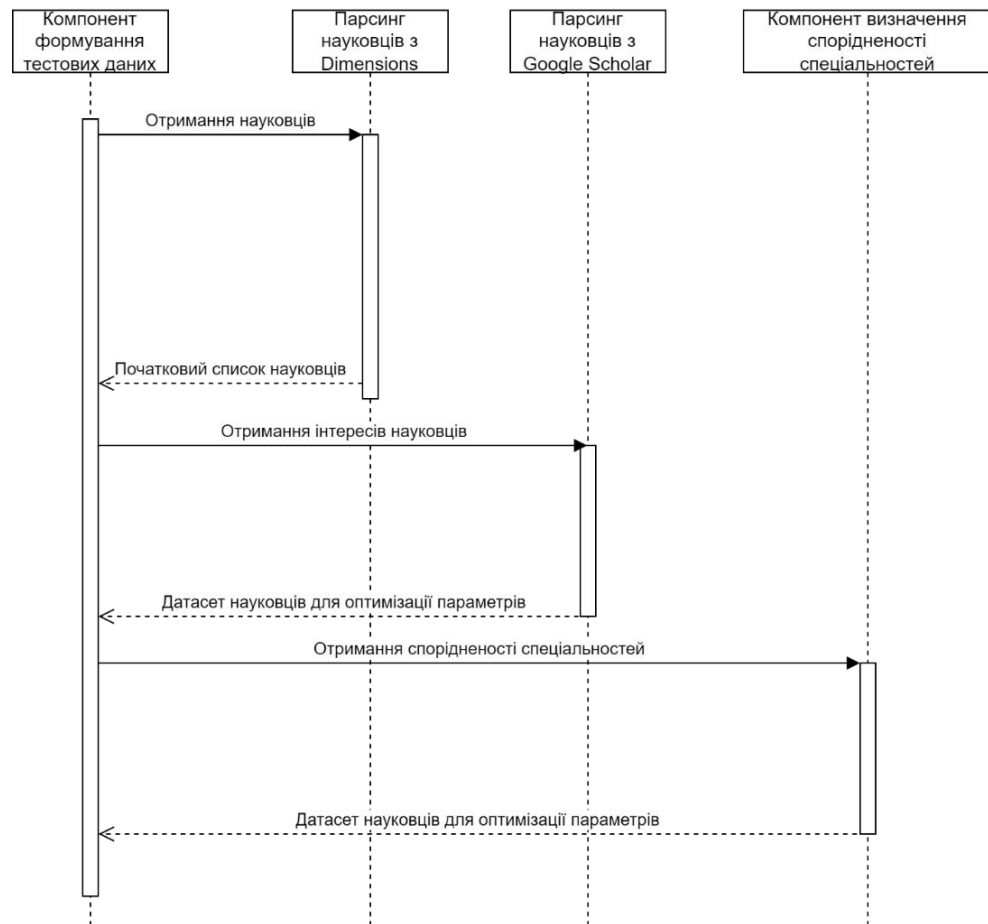


Рисунок 5.6 – UML-діаграма послідовності роботи програмного компоненту формування тестових даних для оптимізації параметрів алгоритму

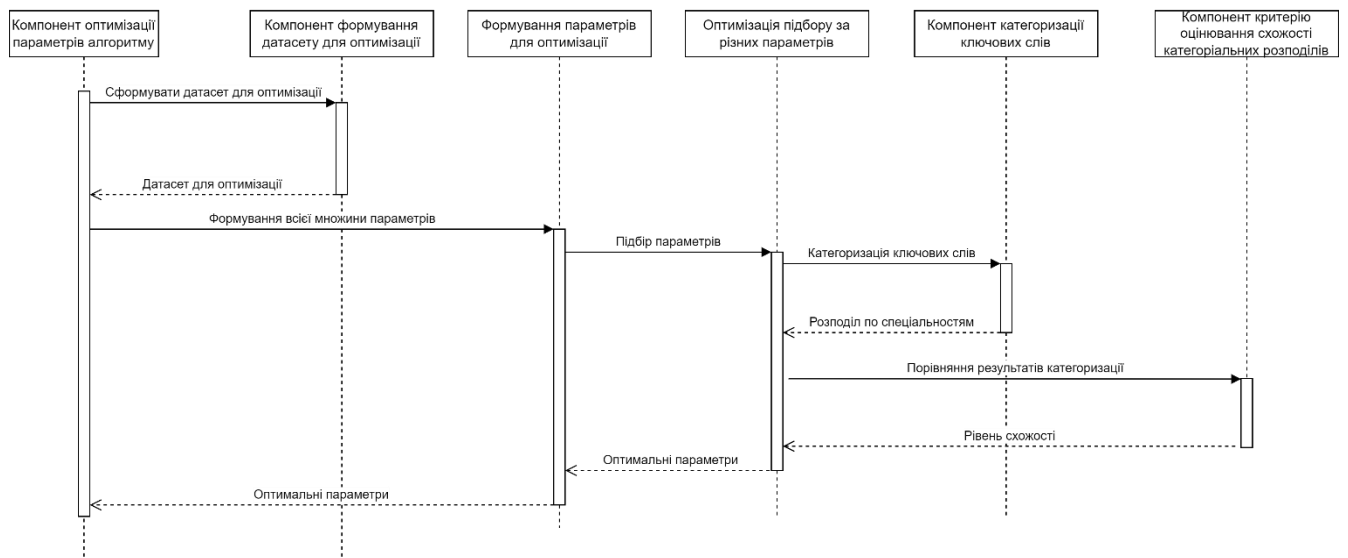


Рисунок 5.7 – UML-діаграма послідовності роботи програмного компоненту оптимізації параметрів алгоритму категоризації ключових слів

5.1.7 Програмний компонент оптимізації призначень опонентів до дисертацій

Програмний компонент призначень опонентів до дисертацій розроблений для ілюстративних цілей, щоб показати як працює запропонований підхід до експрес-підбору рецензентів. Він використовує компонент категоризації ключових слів для категоризації як ключових слів дисертацій, так і ключових слів статей кандидатів. Також він використовує компонент критерію оцінювання схожості категоріальних розподілів. Після категоризації та знаходження схожості відбувається підбір опонентів із застосування різноманітних підходи до оптимізації. Окрім цього реалізовано візуалізацію даних по датасету та аналіз результатів оптимізації оптимізації призначень за різними підходами. UML-діаграма послідовності роботи цього компоненту подана на рис. 5.8.

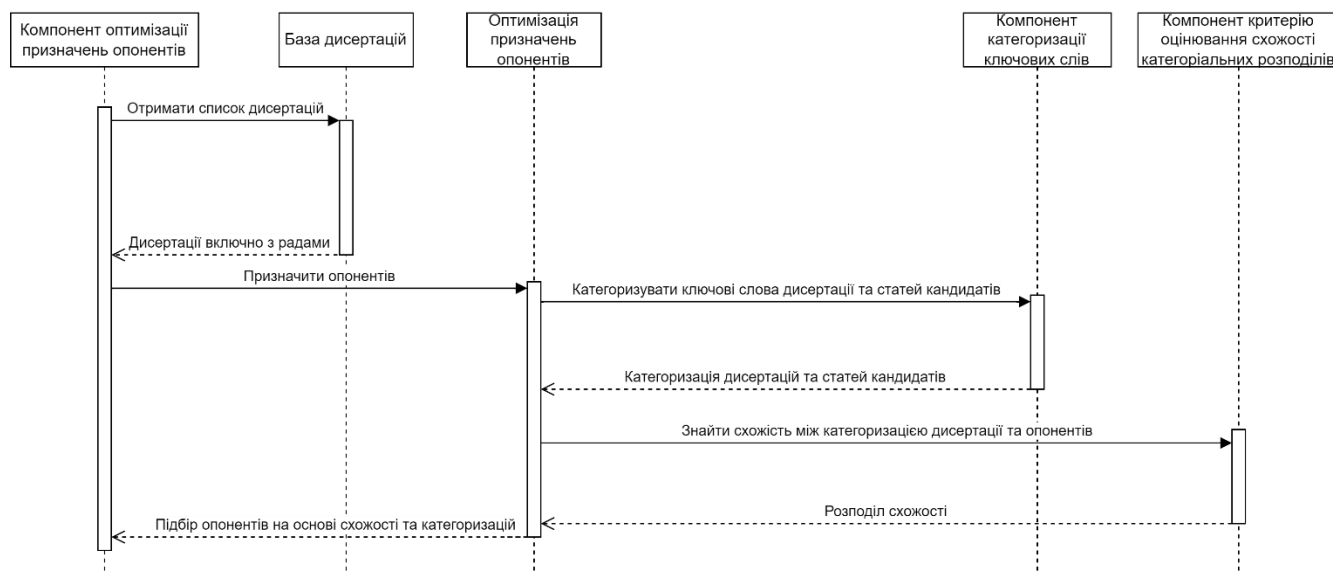


Рисунок 5.8 – UML-діаграма послідовності роботи програмного компоненту оптимізації призначень опонентів до дисертацій

5.1.8 Загальна схема процесу підбору колективу рецензентів

Загальна схема процесу підбору колективу рецензентів з використанням розроблених програмних компонентів наведена на рис. 5.9. З неї видно, що

процес складається з двох головних та двох другорядних етапів. На вході маємо заявку (науковий твір) та базу потенційних рецензентів, яка формується ззовні. База рецензентів складається з ключових слів рецензентів, що вказані ними як їхні наукові інтереси або ж взяті з їх статей. Далі на основі ключових слів заявки та ключових слів кожного рецензента виконується їх категоризація до спеціальностей науки ANZSRC-2020, в результаті якої отримуємо відповідні категоріальні розподіли. Результати категоризації далі використовуються на етапі оптимізації призначення колективу рецензентів до заявки із забезпеченням максимальної відповідності колективу рецензентів до тематики заявки в залежності від того, який критерій обрано – якість підбору чи витрати ресурсів на підбір.

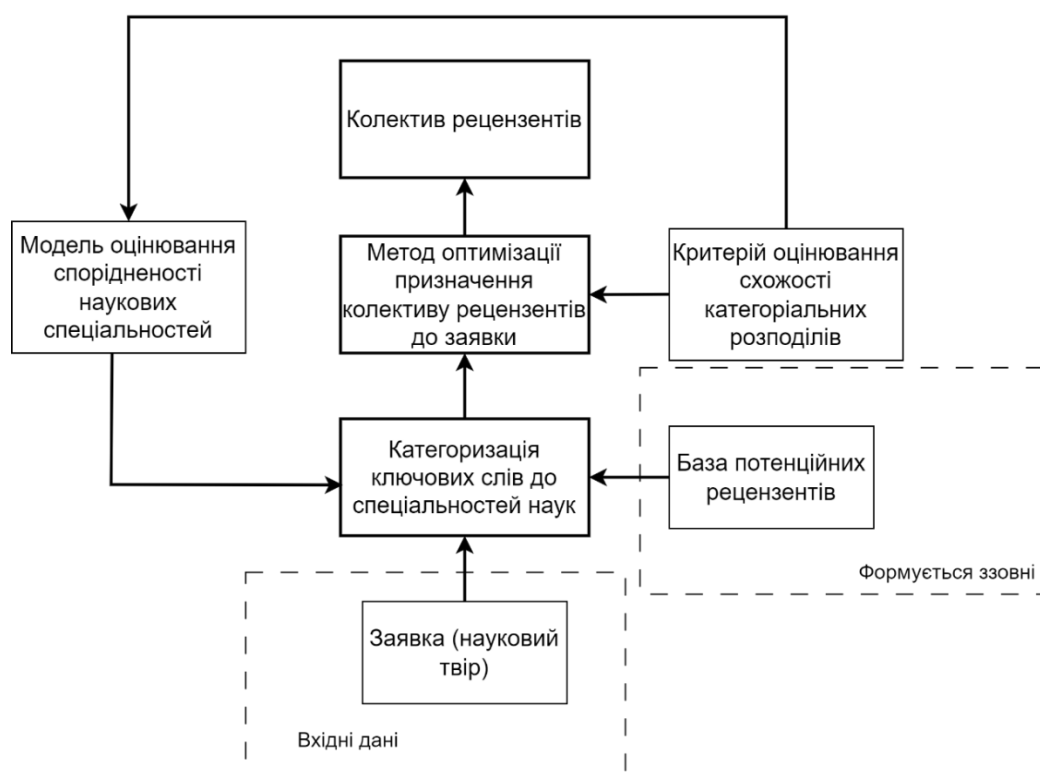


Рисунок 5.9 – Загальна схема процесу підбору колективу рецензентів

5.2 Формування датасету дисертацій

Для проведення експериментів з підбору рецензентів сформуємо датасет дисертацій. Для цього скористаємось інформаційною системою НАЗЯВО

NAQA.Svr [172]. У цій системі подаються заявки на разові спецради для затвердження їх Міністерством освіти і науки України. Кожна дисертація містить перелік супровідних документів, а також дані про разову раду, що пропонується закладом. Разова рада складається з голови, двох або одного рецензента та двох або трьох опонентів відповідно. Відповідність члена ради до тематики дисертації встановлюється на основі статей члена ради.

В результаті зібрано дані по 67 дисертаціям. 17 з них були відхилені міністерством через невідповідність тематик статей членів ради тематиці дисертацій. 50 дисертацій було захищено на момент формування датасету, включаючи 17 повторно сформованих рад, які були раніше відхилені через невідповідність тематиці. Зібрані дисертації відносяться до різних освітніх спеціальностей (рис. 5.9). Значна частина з них відноситься до 12 галузі науки, що включає в себе спеціальності: 121, 122, 123, 124, 125 та 126 (див. Додаток В). В датасет входять: унікальний ідентифікатор дисертації; прізвище та ім'я здобувача; ключові слова дисертації англійською мовою; список членів разової ради разом з ключовими словами їх статей та належністю до дисертації; список проблемних дисертацій (ті які були відхилені через невідповідність тематики). Датасет доступний за посиланням [2].

Загалом було зібрано інформацію за 292 членами різних разових рад. З розподілу кількості ключових слів дисертацій та кількості ключових слів статей членів разових рад (рис. 5.10) видно, що кількість ключових слів для обох випадків знаходиться в межах 40, при чому для дисертацій кількість ключових слів дещо менша з зрозумілих причин. При цьому спостерігаються випадки порушення вимог до дисертацій, а саме кількості ключових слів, яка повинна буде в межах від 5 до 15.

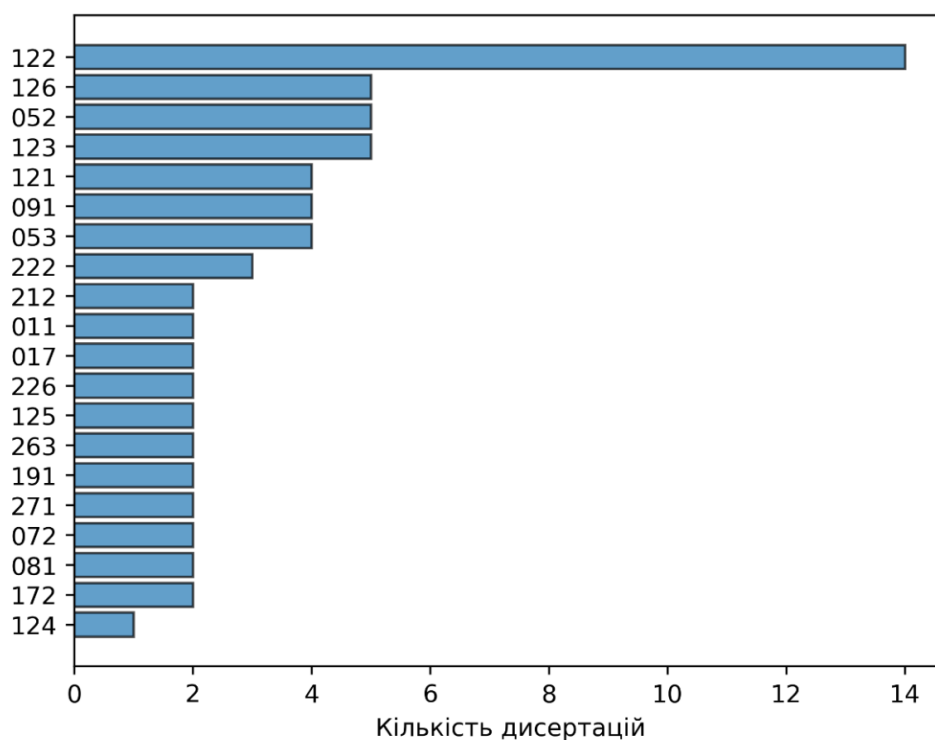


Рисунок 5.9 – Розподіл кількості дисертацій за спеціальностями

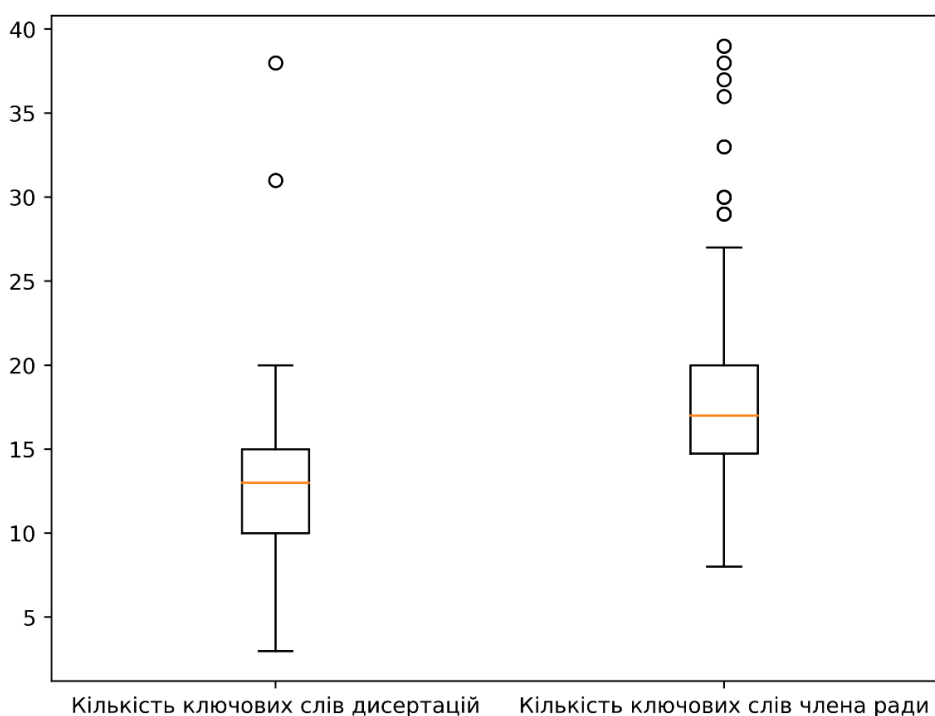


Рисунок 5.10 – Розподіл кількості ключових слів дисертацій та членів разових рад

Маючи сформований датасет, проаналізуємо відповідність членів разових рад тематиці дисертацій використовуючи запропоновані методи представлення даних. Для цього спочатку категоризуємо ключові слова дисертації за запропонованим алгоритмом категоризації ключових слів в межах

спеціальностей наук з ANZSRC-2020. Далі аналогічним чином категоризуємо ключові слова статей членів разових рад з врахуванням попарного поєднання ключових слів лише в межах однієї статті. Після цього визначимо відповідність кожного окремого члена разової ради тематиці дисертацій (незалежно від комплексної відповідності усієї ради дисертації).

Результат визначення схожості подано на рис. 5.11. На ньому по осі абсцис подано ранг членів рад – порядковий номер члена ради у відсортованому списку членів за спаданням схожості з відповідними дисертаціями. На рисунку наведено 2 розподіли: один показує розподіл членів рад, яких затверджено з першого разу, а другий показує розподіл членів разової ради, для тих дисертацій, які були відхилені через невідповідність тематиці (проблемні члени). З розподілів видно, 35 з них зовсім не відповідають тематиці дисертації для тих дисертацій, хоча і не були зупинені міністерством. У проблемних спецрадах 8 членів мають нульову відповідність тематиці дисертації.

Розглянемо детальніше схожість членів разових рад, а саме співвідношення схожості рецензентів (включно з головою) та опонентів. Для цього порахуємо кількість випадків, коли рецензенти більше відповідають тематиці дисертації, ніж опоненти та навпаки. Аналіз проведемо для трьох випадків:

абсолютна перевага – кількість дисертацій, в яких найменш схожий рецензент/опонент більш схожий до дисертації, ніж відповідний найбільш схожий опонент/рецензент;

значна перевага – кількість дисертацій, в яких найбільш схожий рецензент/опонент переважає найбільш схожого опонента/рецензента і найменш схожий рецензент/опонент переважає найменш схожого опонента/рецензента;

помірна перевага – кількість дисертацій в яких середня схожість рецензентів/опонентів вища ніж опонентів/рецензентів. Звичайна перевага – кількість

рецензентів/опонентів що більш схожі до дисертації ніж відповідні опоненти/рецензенти.

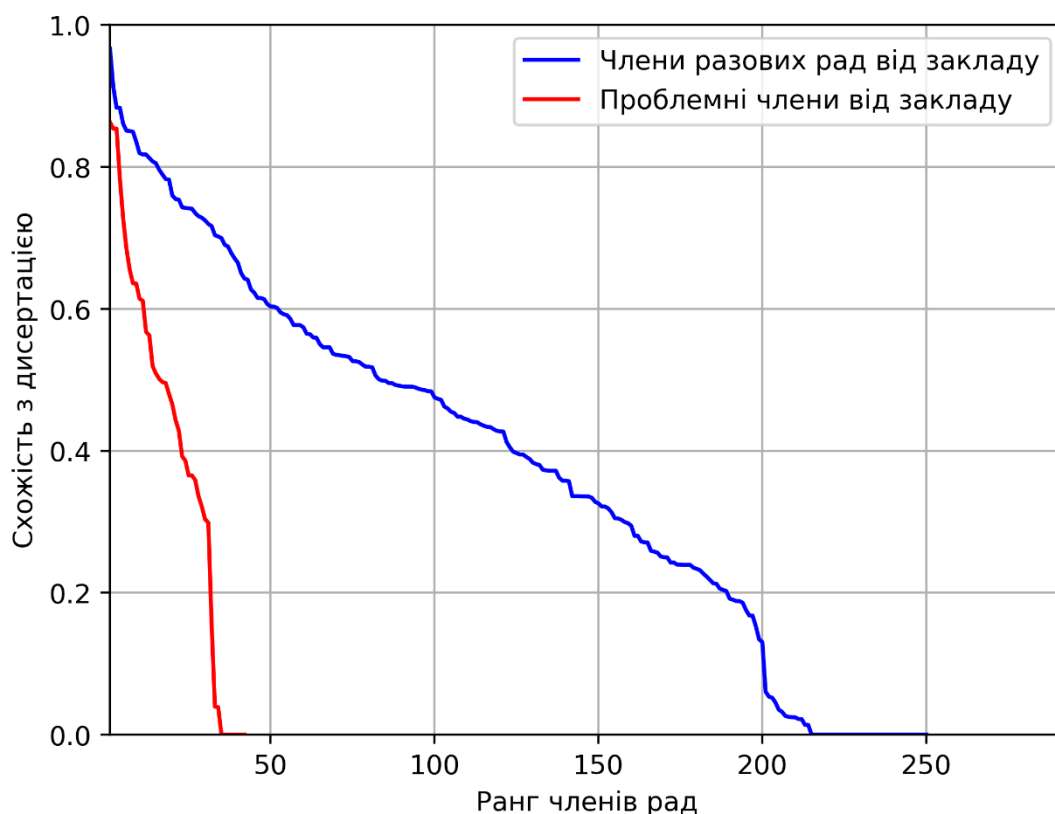


Рисунок 5.11 – Ранговий розподіл відповідності членів ради тематиці дисертації

Результати порівняння зведено у табл. 5.1. Видно, що розподіл переваг приблизно однаковий на всьому датасеті дисертацій, і це при тому, що опонентів зазвичай менше.

Таблиця 5.1 – Порівняння відповідності рецензентів та опонентів

Рівень переваги	Кількість рецензентів	Кількість опонентів
Абсолютна перевага	11	9
Значна перевага	17	10
Помірна середнього	22	25

5.3 Підбір опонентів

Для проведення експериментів розглянемо задачу підбору опонентів на сформованому датасеті дисертацій. Для кожної дисертації вилучимо опонентів заявлених від закладу, і підберемо тих які найкраще підходять до дисертації за

певними критеріями. Після вилучення опонентів отримуємо множину разових рад, що містять лише голову та двох рецензентів (або одного рецензента в залежності від початкового складу). Як базу потенційних опонентів використаємо членів разових рад інших дисертацій.

Більш формально маємо множину дисертацій з розподілом ключових слів за спеціальностями та розподілом ключових слів статей членів разової ради від закладу без вилучених опонентів:

$$H = \left\{ \left(W_1^H, \{W_1^R, \dots, W_K^R\} \right), \dots, \left(W_N^H, \{W_1^R, \dots, W_K^R\} \right) \right\}, \quad (5.1)$$

де $K \in \{2, 3\}$ - кількість членів ради без опонентів, $i = \overline{1, N}$.

Необхідно знайти опонентів, що забезпечують найбільшу схожість разової ради з дисертацією:

$$H^* = \left\{ \left(W_i^H, \{W_1^R, \dots, W_C^R\} \in W^R \right) \middle| \max \left(\text{Fit} \left(W_i^H, \text{Aggregate} \left(\{W_1^R, \dots, W_C^R\} \right) \right) \right) \right\}, \quad (5.2)$$

де $C = 5$ - кількість членів разової ради;

$\text{Aggregate}(\{W_1^R, \dots, W_C^R\})$ - функція агрегації розподілів членів потенційної разової ради за спеціальностями;

$\text{Fit}(W_i^H, \text{Aggregate}(\{W_1^R, \dots, W_C^R\}))$ - рівень відповідності разової ради тематиці дисертації за критерієм (3.1).

На рис. 5.12 подано ранговий розподіл відповідності разової ради запропонованої закладом до дисертації. З рисунка видно, що більша частина разових рад мають відповідність вище 0.2 (62 ради), а решта мають незначущий рівень відповідності (4 ради). Міжквартильний інтервал знаходиться приблизно в межах $[0.4; 0.8]$.

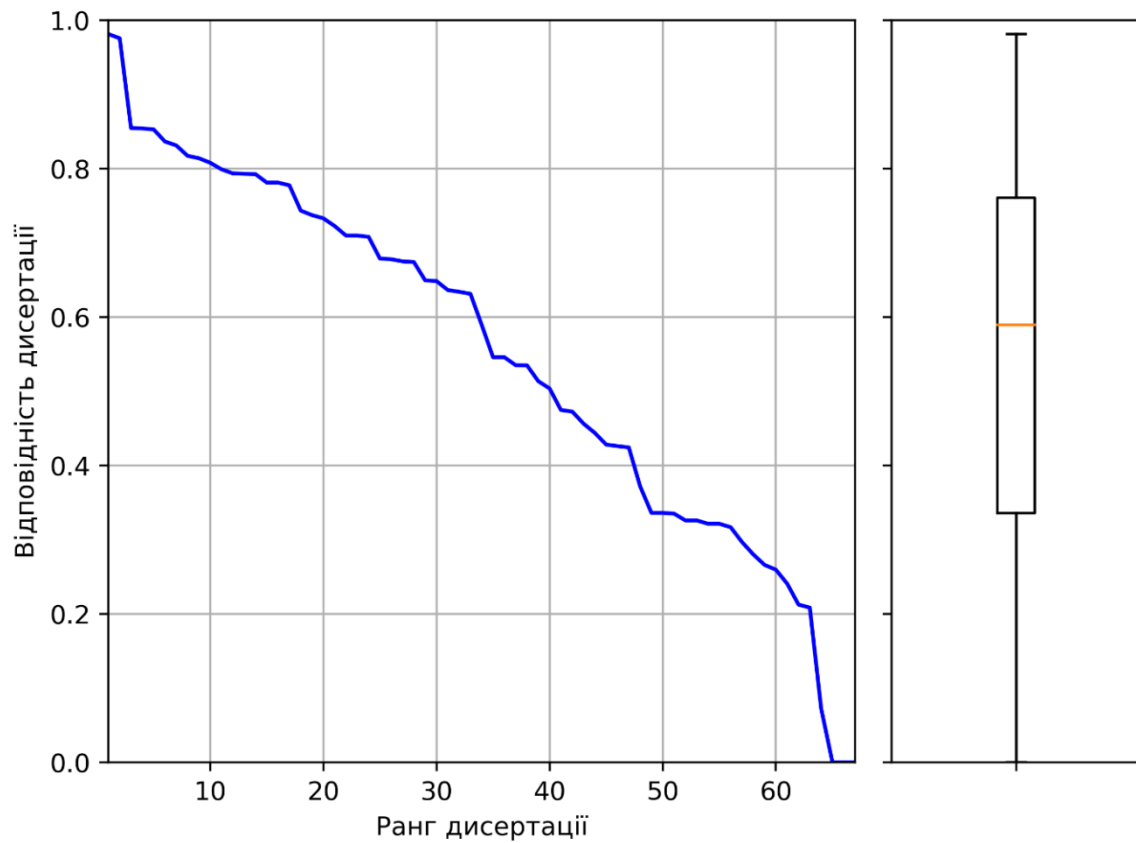


Рисунок 5.12 – Ранговий розподіл відповідності дисертації разовій раді, сформованій закладом

5.3.1 Оцінка ефекту від оптимізації

Результати підбору опонентів порівнюємо з базовим варіантом складу разової ради, який сформований закладом. Кількісно ефект оцінемо середнім рівнем зміни відповідності разових рад. Відповідність разової ради закладу визначається наступним чином:

$$F = \left\{ \text{Fit}(W_i^H, \text{Aggregate}(\{W_1^R, \dots, W_C^R\})) \mid i = \overline{1, N} \right\}$$

Аналогічним чином відповідність запропонованої разової ради закладу має вигляд:

$$F^* = \left\{ \text{Fit}(W_i^{H^*}, \text{Aggregate}(\{W_1^R, \dots, W_C^R\})) \mid i = \overline{1, N} \right\}.$$

Тоді середнє покращення має вигляд:

$$I(F^*, F) = \frac{\sum_{i=1, N} (F_i^* - F_i)}{N} \cdot 100.$$

5.3.2 Оптимізація за повним перебором

Повний перебір дає оптимальний результат, проте для великих обсягів даних перебрати всі можливі варіанти та вкласти в практичні часові обмеження нереально. Приміром, на тестовому датасеті за врахування лише потенційних опонентів з не нульовою схожістю з дисертацією така оптимізація зайняла 7 годин. Тому оптимізація за повним перебором має на меті ідентифікувати межі покращення якості підбору, щоб можна було їх використовувати як реперну точку для оцінювання інших більш практичних алгоритмів оптимізації.

На рис. 5.13 подано результати оптимізації за повним перебором. Середнє покращення становить 55%. Можна помітити суттєву різницю між відповідністю ради від закладу та запропонованою радою. Видно, що разові ради від закладу мають значно нижчу відповідність тематиці дисертацій, ніж знайдені за повним перебором. 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 за достатнього великого пулу кандидатів отримуємо значне покращення рад лише за рахунок підбору опонентів.

На рис. 5.14 подано точкову діаграму порівняння відповідностей за кожною дисертацією. Спостерігається значне покращення відповідності рад, як тих які були відхилені, так і всіх інших. Можна помітити, що деякі ради не були покращені або рівень покращення низький. Це зумовлено в першу чергу тим, що в датасеті розподіл дисертацій за спеціальностями є нерівномірним (див. рис. 5.9) та малим обсягом датасету.

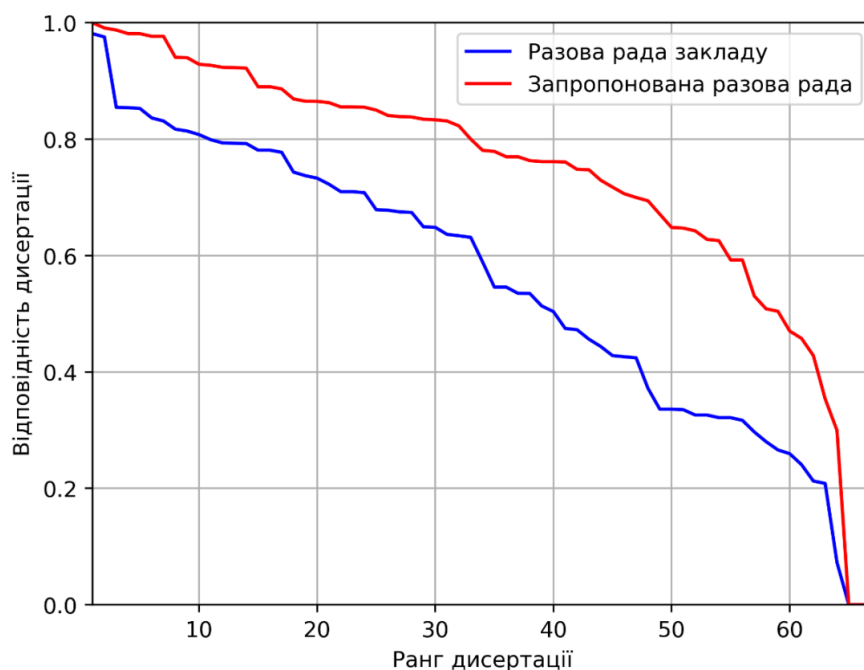


Рисунок 5.13 – Порівняння розподілу відповідності разових рад від закладу з запропонованими радами

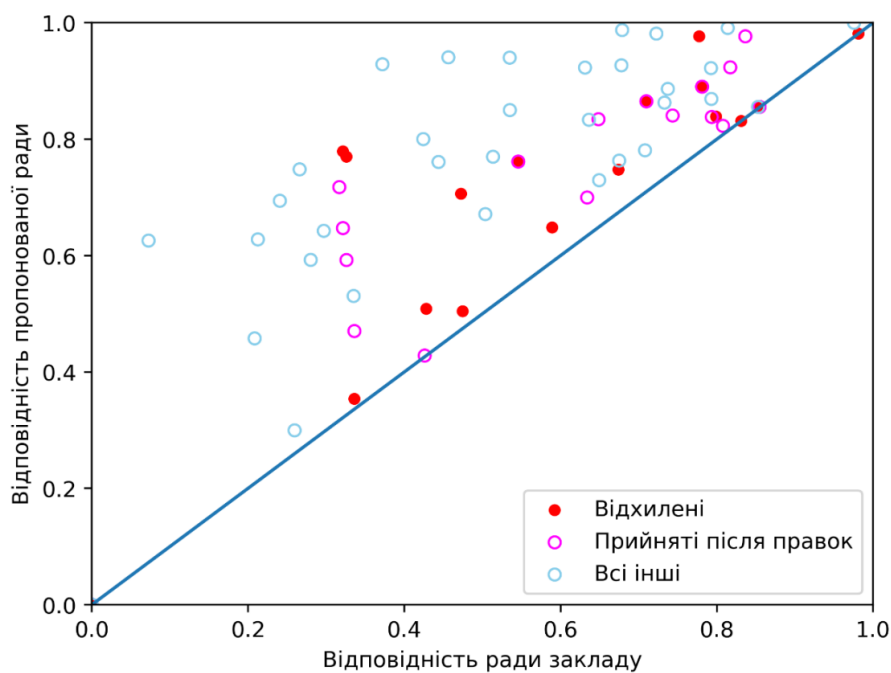


Рисунок 5.14 – Точкова діаграма порівняння відповідностей рад від закладу з пропонованими радами

Повний перебір враховує всі можливі комбінації пар або трійок опонентів. Хоча на практиці опоненти з низьким рівнем відповідності до дисертації наврядчи будуть призначені до неї. В зв'язку з цим виникає питання

визначення балансу «рівень покращення – тривалість перебору». Дослідимо, як зміниться середній рівень покращення, якщо знехтувати деякими потенційними опонентами з низькою схожістю і таким чином зменшити час перебору. Для цього проведемо 3 обчислювальні експерименти за повним перебором на обмеженому списку кандидатів – попередньо відкинемо потенційних опонентів з рівнем схожості менше ніж 0.1, 0.2 та 0.3. Результати цих трьох експериментів подано на рис. 5.15. На ньому зображено 5 кривих: крива розподілу відповідності разових рад від закладу, крива розподілу відповідності разових рад утворених за повним перебором, та 3 кривих розподілів разових рад утворених повним перебором з врахуванням лише кандидатів з рівнем схожості вище або дорівнює 0.1, 0.2 та 0.3 відповідно.

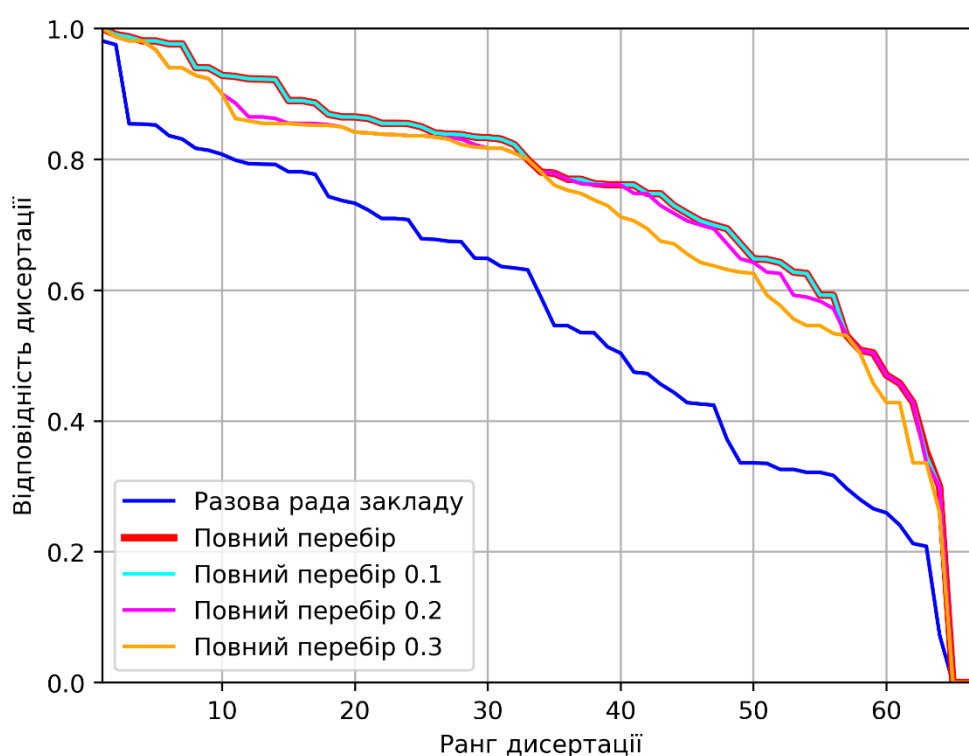


Рисунок 5.15 – Порівняння результатів повного перебору за різних порогових обмежень на відповідність кандидата дисертації

З рис. 5.15 видно, нехтуванням кандидатами з низькою відповідністю некритично знижує результати оптимізації. При цьому тривалість перебору суттєво знижується. За порогового обмеження у 0.1, тривалість оптимізації

скоротилася у 2.6 рази, а середнє покращення залишилось майже на тому самому ж рівні. За порогового обмеження у 0.1 час скоротився у 5 разів, а середнє покращення зменшилось лише на 5%. Це підтверджує той факт, що при достатній кількості наявних кандидатів з високим рівнем схожості до дисертацій можна нехтувати слабкими кандидатами для пришвидшення роботи повного перебору.

5.3.3 Оптимізація за жадібним алгоритмом

На рис. 5.16 подано порівняння результатів підбору ради за жадібним алгоритмом та за повним перебором. Результати за жадібним алгоритмом трохи гірше, проте різниця не велика. Середнє покращення за жадібним алгоритмом становить 51.5%, а для повного перебору – 55.2%. Проте час підбору опонентів зменшився до 493 секунд, що в 49 разів швидше за повний перебір.

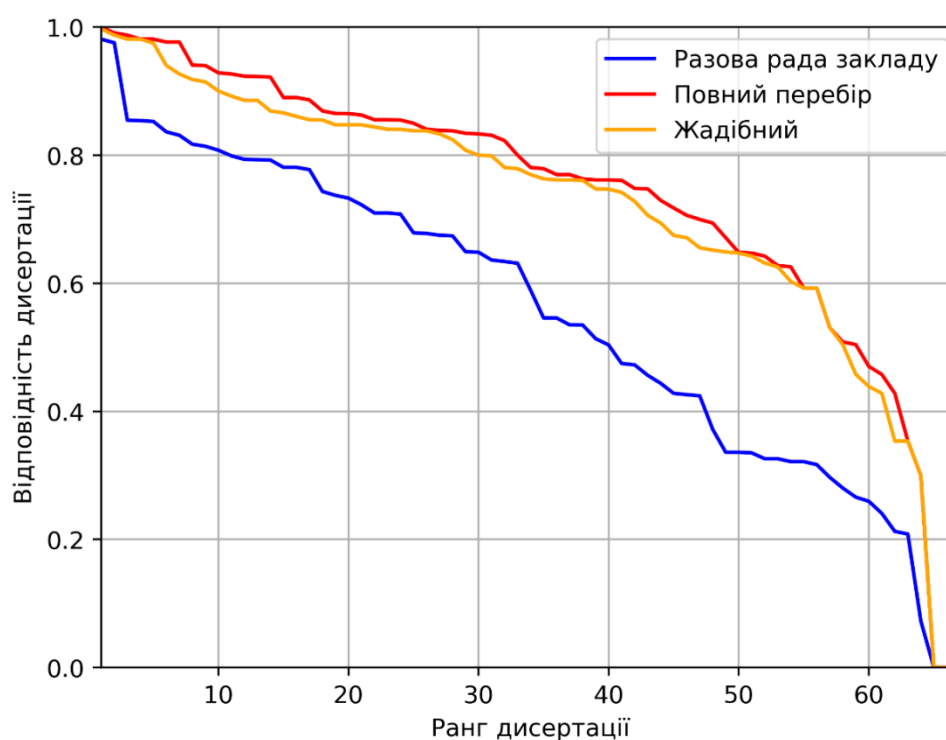


Рисунок 5.16 – Порівняння результатів оптимізації за жадібним алгоритмом та за повним перебором

На рис. 5.17 подано точкову діаграму порівняння разових рад від закладу з пропонованими радами, які сформовано за жадібним алгоритмом. Оскільки

жадібний алгоритм не гарантує оптимального розв'язку то можна спостерігати, що декілька запропонованих рад мають гіршу відповідність ніж ради від закладу. Але таких рад дуже мало.

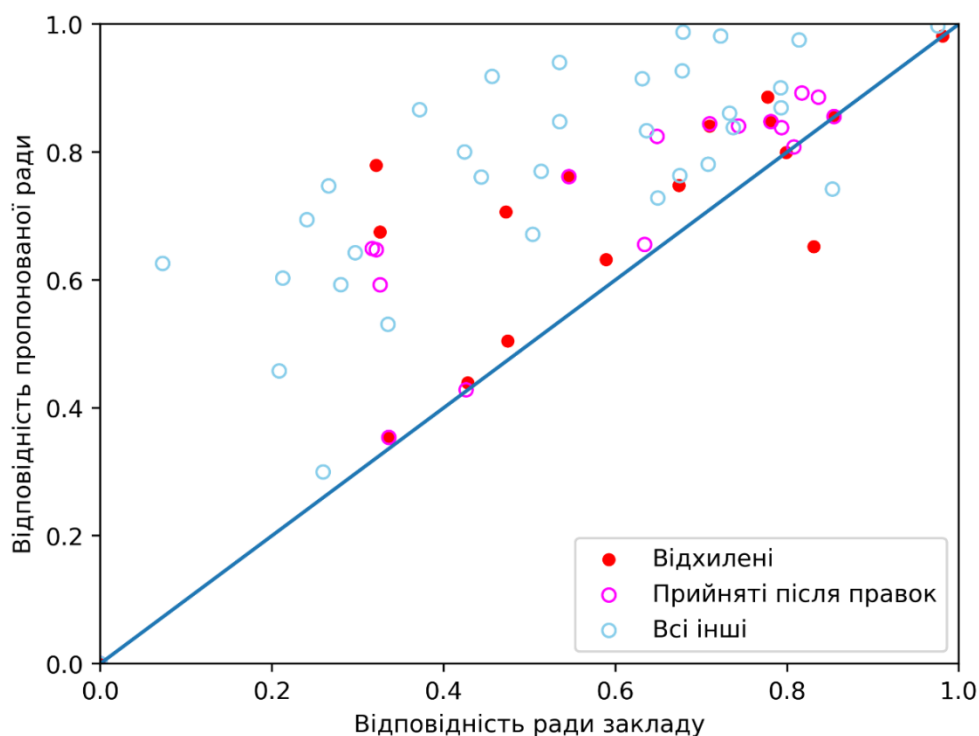


Рисунок 5.17 – Порівняння відповідностей рад від закладу з запропонованими радами, які підібрано за жадібним алгоритмом

5.3.4 Оптимізація за жадібним алгоритмом з елітизмом

Інша варіація жадібного алгоритму це підбір за принципом елітизму. У цьому випадку перший опонент підбирається з найбільшим рівнем схожості з дисертацією (еліта), а інший (або два інших) підбирається класичним жадібним алгоритмом. Такий підхід зменшує кількість операцій, але при цьому і якість підбору теж буде гіршою. На рис. 5.18 порівнюються результати оптимізації за жадібними алгоритмами та повним перебором. Середнє покращення для жадібного алгоритму з елітизмом становить 47.7%, що гірше ніж для інших аналізованих алгоритмів. Проте час підбору зменшився майже вдвічі в порівнянні з класичним жадібним та в 97 разів у порівнянні з повним перебором.

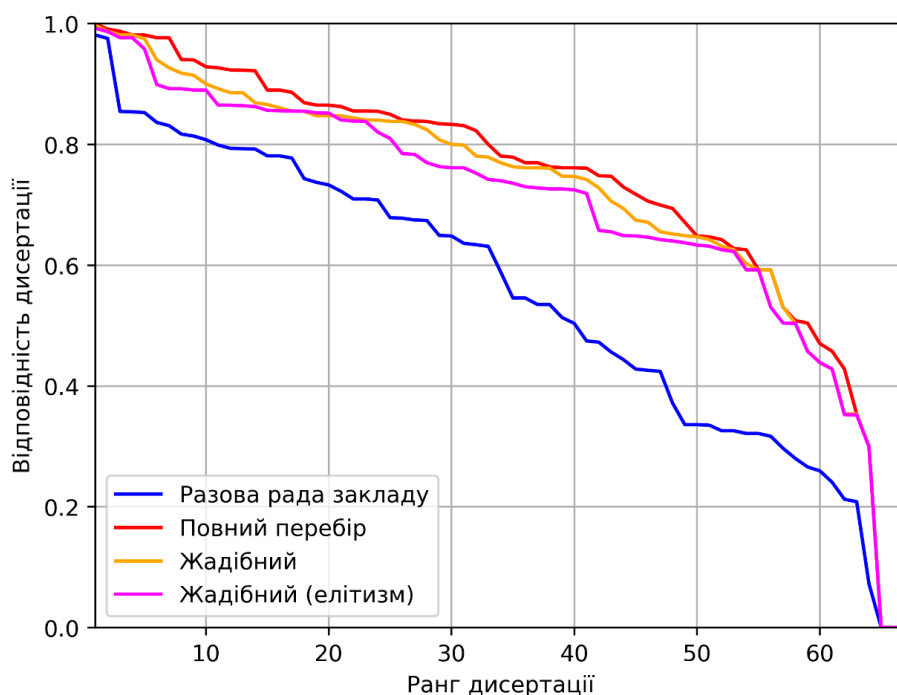


Рисунок 5.18 – Порівняння результатів підбору рад за жадібними алгоритмами

На рис. 5.19 подано точкову діаграму порівняння разових рад закладу з пропонуваними радами, які підбрано за жадібним алгоритмом з елітизмом. У порівнянні з жадібним алгоритмом та повним перебором можна спостерігати, що значна частина запропонованих рад мають гіршу відповідність проте все ще значна частина краща за ради від закладу.

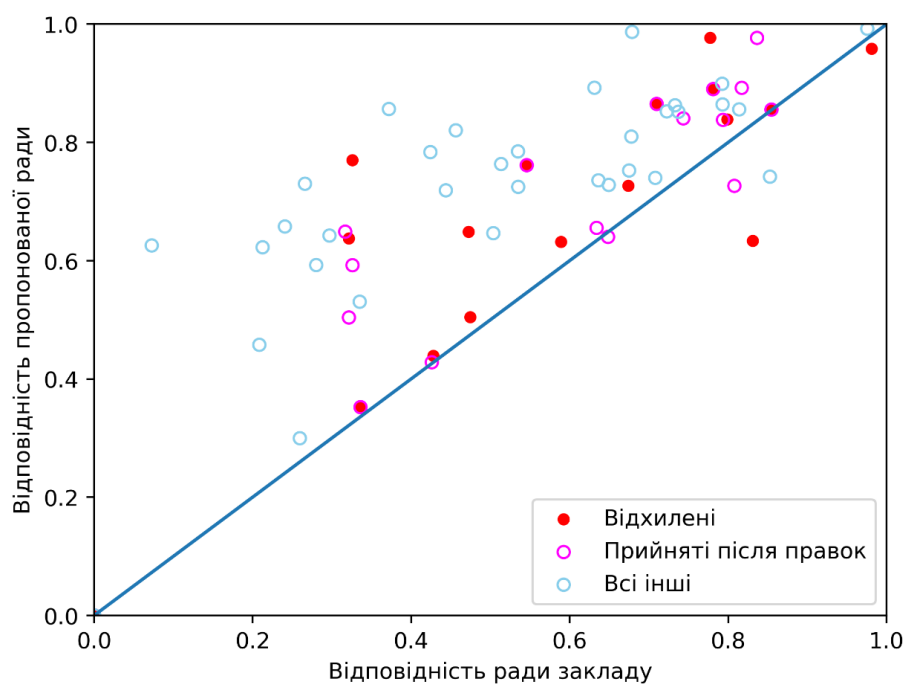


Рисунок 5.19 – Порівняння відповідностей рад від закладу з радами, які сформовано за жадібним алгоритмом з елітизмом

5.3.5 Оптимізація за найбільш схожими кандидатами

Найпростіший спосіб призначити рецензентів це обрати тих, що найбільш схожі до тематики дисертації. На рис. 5.20 подано порівняння результатів роботи цього підходу. З графіка видно, що розподіл приблизно вдвічі гірший за повний перебір. Середнє покращення становить 32.3%, у порівнянні з повним перебором (55.2%) це в 1.7 разів гірше. Проте час підбору на тестовому датасеті зменшився до 7 секунд, що значно швидше за всі інші підходи. На рис. 5.21 подано точкову діаграму для порівняння разових рад закладу з пропонованими радами, які підібрано за найбільш схожими кандидатами. Спостерігається, що значна кількість запропонованих рад мають гіршу відповідність ніж ради від закладу, це пояснюється в першу чергу тим що висока схожість з дисертацією не означає, що утворений колектив буде покривати усю тематику дисертації.

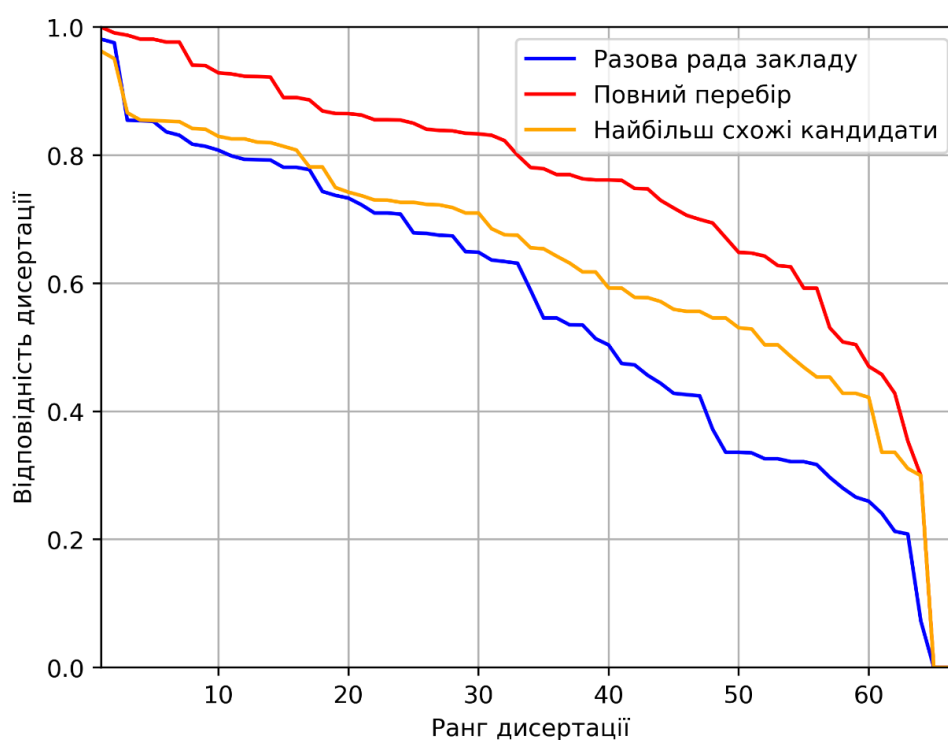


Рисунок 5.20 – Порівняння результатів оптимізації за найбільш схожими кандидатами та за повним перебором

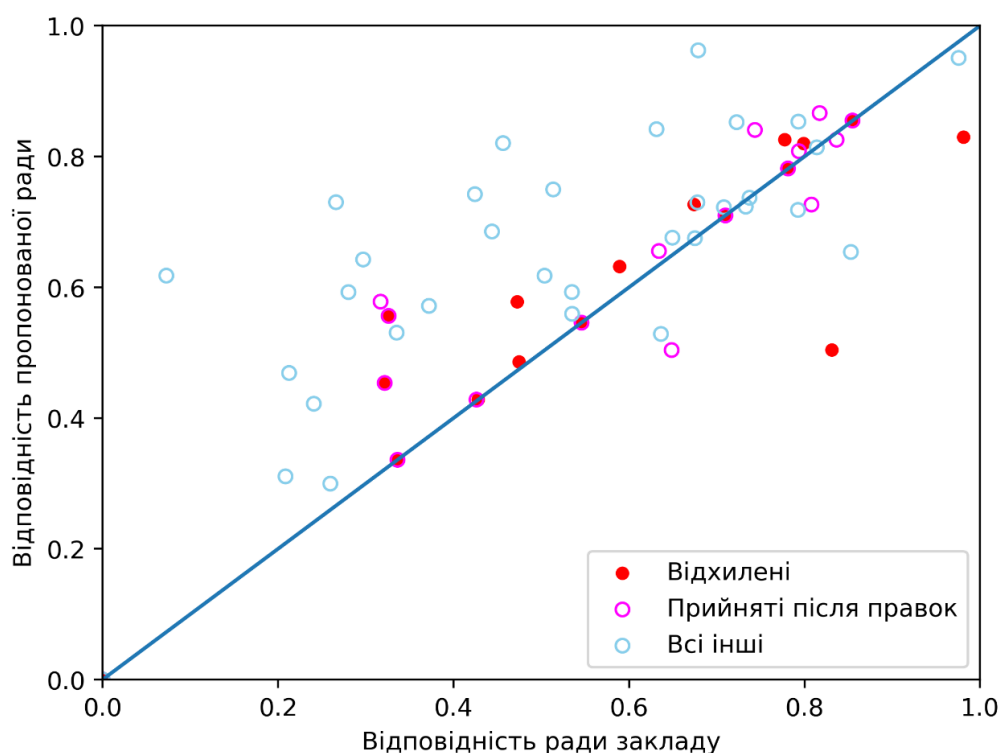


Рисунок 5.21 – Порівняння рад від закладу з пропонованими радами, які підібрано за найбільш схожими кандидатами

5.3.6 Аналіз результатів оптимізації

В табл. 5.2 наведено порівняння часу оптимізації та середнього покращення. Результати оптимізації підбору опонентів за різних підходів також подано на рис. 5.22. З рисунку видно, що перебір на обрізаній множині з пороговим обмеженням у 0.3 однозначно невдалий. Усі інші утворюють множину Парето. Тому під час виборі алгоритму необхідно враховувати пріоритети, чи потрібен швидкий результат чи якісний.

Таблиця 5.2 – Характеристики часу виконання та середнього покращення різних варіантів оптимізації

Алгоритм підбору	Час виконання, секунди	Середнє покращення $I(F^*, F)$, %
Повний перебір	24251	55.2
Повний перебір 0.1	9052	55.1
Повний перебір 0.2	4767	52.5
Повний перебір 0.3	1835	48.6
Жадібний	493	51.5
Жадібний з елітизмом	250	47.7
Максимальна схожість	7	32.3

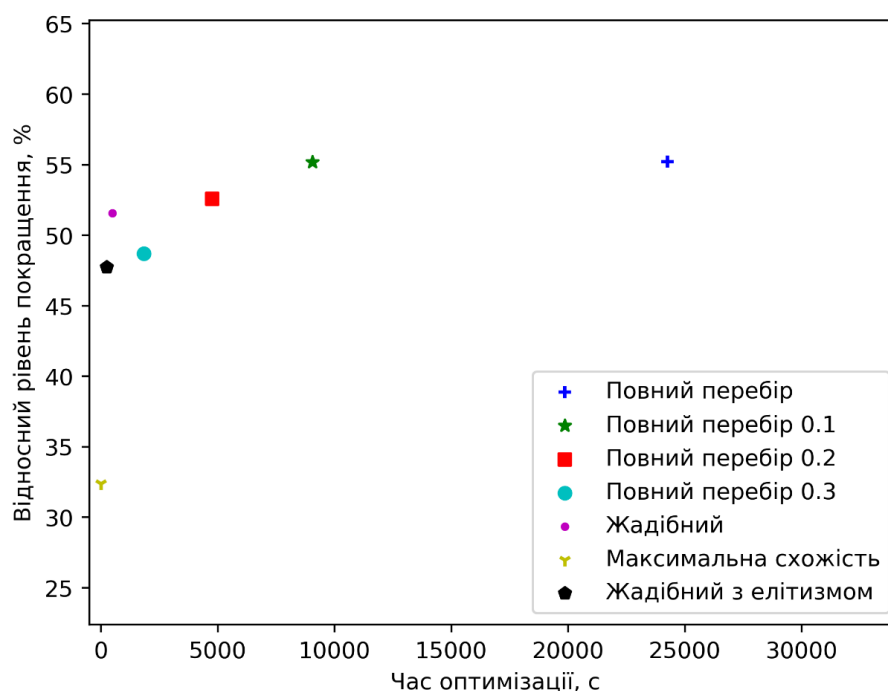


Рисунок 5.22 – Співвідношення «тривалість – якість» різних алгоритмів підбору опонентів

5.3.7 Приклад покращення разової ради від закладу

Розглянемо приклад того, як за допомогою повного перебору вдалося покращити склад разової ради в порівнянні з складом пропонованим закладом. Для цього покроково покажемо результати підбору для такої дисертації.

Автор дисертації: *Левківський Віталій Леонідович*.

Назва дисертації: *Моделі та методи обробки даних системи віддаленого моніторингу стану пацієнтів з цукровим діабетом*.

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*.

Результат категоризації ключових слів дисертації:

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.

$$\text{Тобто, } P^* = \left(\frac{0.382}{4605}, \frac{0.255}{4606}, \frac{0.205}{4609}, \frac{0.158}{4203} \right).$$

Члени разової ради від закладу.

1. Голова разової ради.

Ключові слова статей: *a) information technology, cardiovascular system, diagnostics, SCORE, Systematic Coronary Risk Evaluation, blood pressure, diseases, medical information systems; б) information technology, data processing, data, data visualization, COVID-19, mobile application, disease, medical information systems; в) Differential equations, Acceleration, Cloud computing, Task analysis, Computational complexity, Computer simulation, computerized integration procedure, integration procedure, numerical method.*

Результат категоризації:

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.

2. Рецензент 1

Ключові слова статей: *a) IoT, monitoring system, microclimate parameters, educational institutions; б) 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; в) pandemic, health-saving educational environment, model, information and digital environment; г) planning technologies, use case, project, systems design, project complexity.*

Результат категоризації:

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.

3. Рецензент 2

Ключові слова статей: а) *mathematical model, biotechnical system, edge devices, methods of studying, human body health, hemodynamic characteristics, diagnostic systems, digital signal processing, information technology*; б) *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*; в) *IoT, monitoring system, microclimate parameters, educational institutions, edge devices*.

Результат категоризації:

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.

4. Опонент 1

Ключові слова статей: а) *information expert system, control, method of fuzzy sets, medical diagnostics, diabetes, dentistry*; б) *medical information technologies, medical information systems, coronary channels, coronary artery disease*; в) *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*; г) *medical expert systems, fuzzy logic, coronary artery disease, coronary arteries, problems of cardiology, cardiovascular diseases, myocardial infarction, patient safety*.

Результат категоризації:

3201 Cardiovascular Medicine and Haematology – 0.387;

3203 Dentistry – 0.215;

4605 Data Management and Data Science – 0.205;

4602 Artificial Intelligence – 0.192.

5. ОПОНЕНТ 2.

Ключові слова статей: а) *Data analysis, Computational modeling, Neural networks, Predictive models, Safety, Personnel, Task analysis, hypertension, medical robotic platforms, TIMA, health and safety digital competencies*; б) *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*; в) *dermatoscopy, medical diagnosis, convolutional neural network, skin disease, ResNet50 model, software component model*.

Результат категоризації:

4602 Artificial Intelligence – 0.435;

4611 Machine Learning – 0.357;

4605 Data Management and Data Science – 0.208.

Для наочності зведемо всі дані до табл. 5.3. Бачимо, що всі члени разової ради мають значний рівень схожості з дисертацією. Далі необхідно знайти рівень відповідності разової ради до дисертації. Для цього необхідно агрегувати результати категоризації усіх членів разової ради. Після агрегування отримаємо:

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

Знайдемо відповідність разової ради від закладу до дисертації, маємо:

$$Fit\left(\left(\frac{0.382}{4605}, \frac{0.255}{4606}, \frac{0.205}{4609}, \frac{0.158}{4203}\right), \left(\frac{0.389}{4606}, \frac{0.374}{4605}, \frac{0.236}{4602}\right)\right) = 0.631. \quad \text{Це досить}$$

непоганий рівень відповідності, можна помітити співпадіння в двох спеціальностях: 4606 *Distributed Computing and Systems Software*, 4605 *Data Management and Data Science*.

Таблиця 5.3 – Результати категоризації профілів членів разової ради від закладу

Член разової ради	Ступінь належності профілю члена до спеціальностей, μ_i	Схожість між дисертацією та членом спецради $Fit(P^*, \mu_i)$
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

Спробуємо підібрати кращих опонентів, яку сукупно будуть підвищувати відповідність разової ради дисертації. Як потенційних кандидатів візьмемо членів усіх інших разових рад сформованого датасету.

Використовуючи повний перебір отримуємо наступних опонентів:

1. Новий опонент 1 – Мелеховець Оксана

Ключові слова статей: а) *trophic ulcers, diabetes mellitus, chronic venous insufficiency, photodynamic therapy, plasma therapy*; б) *diabetes mellitus, diabetic foot, photodynamic therapy, autologous plasma*; в) *type 2 diabetes, middle age, obesity, quality of life, physical therapy program*; г) *type 2 diabetes, middle age, obesity, physical therapy program*.

Результат категоризації:

3210 *Nutrition and Dietetics* – 0.274;

4203 *Health Services and Systems* – 0.261;

4202 *Epidemiology* – 0.251;

3205 *Medical Biochemistry and Metabolomics* – 0.214.

2. Новий опонент 2 - Сторчак Каміла

Ключові слова статей: а) *intelligent data analysis, data mining system, neural network*; б) *web system, machine learning, web development, Heroku, Streamlit*; в) *machine learning methods, neural network, smart home, failures, prediction, data analysis, decision theory, information technology*.

Результат категоризації:

4605 *Data Management and Data Science* – 0.555;

4611 *Machine Learning* – 0.306;

4609 *Information Systems* – 0.139.

В табл. 5.4 подано результати категоризації пропонованих опонентів. Видно, що схожість з дисертацією дещо збільшилась в порівнянні з попередніми опонентами. Замінімо опонентів від закладу пропонованими, в результаті агрегування розподілів по членах нової ради отримаємо:

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

Відповідність нової разової ради з підібраними опонентами до дисертації становить: $\text{Fit} \left(\left(\frac{0.382}{4605}, \frac{0.255}{4606}, \frac{0.205}{4609}, \frac{0.158}{4203} \right), \left(\frac{0.361}{4605}, \frac{0.335}{4606}, \frac{0.156}{4609}, \frac{0.148}{4203} \right) \right) = 0.923$.

В порівнянні з разовою радою від закладу бачимо суттєве покращення рівня відповідності, нова рада тепер містить тіж самі спеціальності що і дисертація. Покращення становить приблизно 46%. Таким чином бачимо, що індивідуальна схожість окремого члену ради може бути посередньою, проте сукупно рівень

відповідності самої ради є високим. Це відбувається за рахунок того, 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. 5.23).

Таблиця 5.4 – Результати категоризації пропонованих опонентів

Член разової ради	Ступінь належності профілю члена до спеціальностей, μ_i	Схожість між дисертацією та опонентом $Fit(P^*, \mu_i)$
Новий опонент 1	$\left(\frac{0.274}{3210}, \frac{0.261}{4203}, \frac{0.251}{4202}, \frac{0.214}{3205} \right)$	0.158
Новий опонент 2	$\left(\frac{0.555}{4605}, \frac{0.306}{4611}, \frac{0.139}{4609} \right)$	0.542

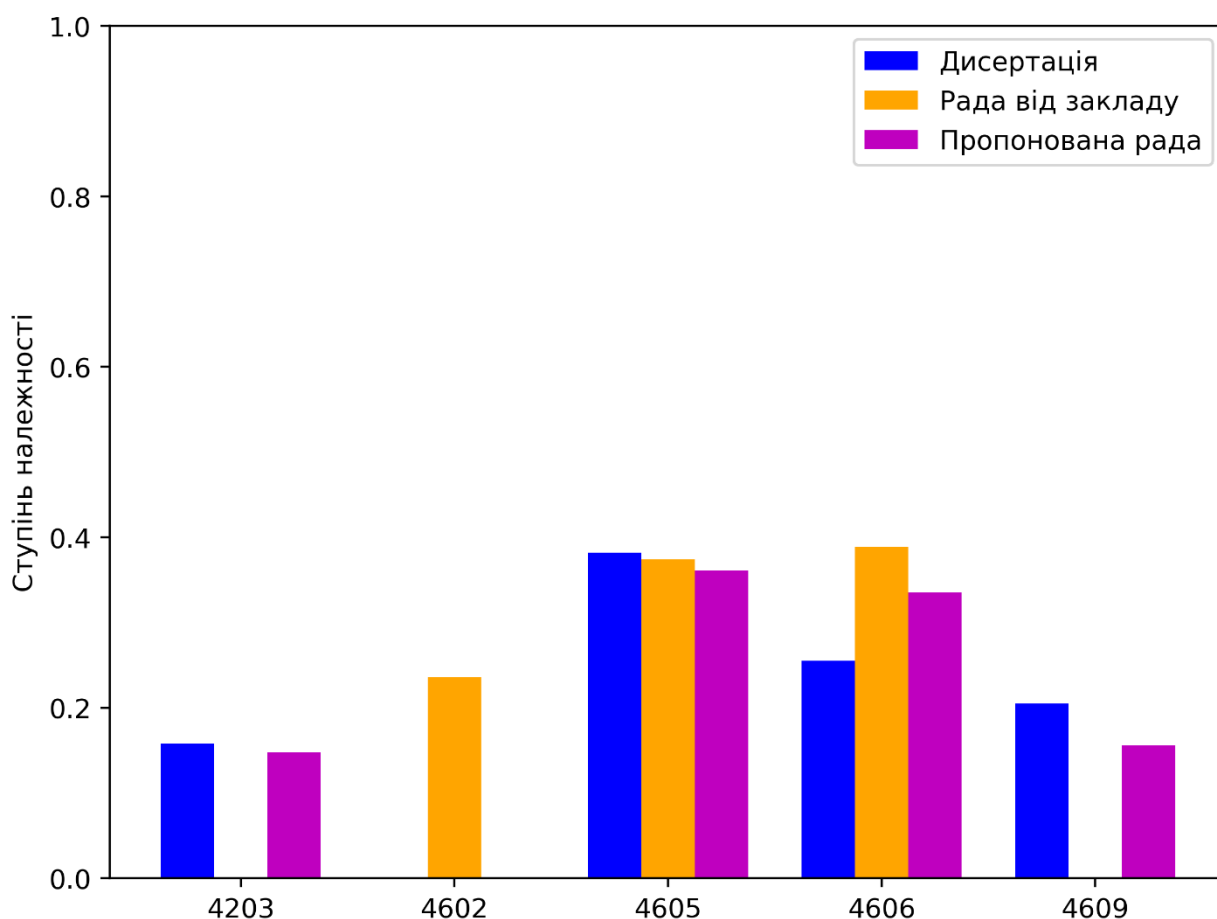


Рисунок 5.23 – Порівняння категоризації ключових слів дисертації з розподілами ради від закладу та пропонованої ради

Висновки до розділу 5

1. Описано високорівневу архітектуру взаємодії між розробленими програмними компонентами інформаційної технології підбору рецензентів для експертизи наукових творів. Розроблено програмний компонент парсингу даних розподілів ключових слів до спеціальностей науки з ANZSRC-2020. Описано роботу програмного компонента категоризації ключових слів, подано блок-схему алгоритму категоризації та UML-діаграму послідовності роботи. Описано програмний компонент критерію оцінювання схожості категоріальних розподілів та розроблено UML-діаграму послідовності роботи компонента. Описано програмний компонент формування тестових даних для оптимізації параметрів алгоритму категоризації ключових слів та розроблено UML-діаграму послідовності роботи компонента. Описано програмний компонент оптимізації параметрів алгоритму категоризації ключових слів та розроблено UML-діаграму послідовності роботи компонента. Описано програмний компонент оптимізації призначень опонентів до дисертацій та розроблено UML-діаграму послідовності роботи компонента. Описано загальну схему роботи процесу підбору колективу рецензентів з детальним описом роботи самої технології підбору.

2. Для проведення експериментів підбору рецензентів сформовано датасет Phd-дисертацій взятих з НАЗЯВО. Значна частина з них відноситься до 12 галузі науки. Проаналізовано відповідність членів рад дисертаціям незалежно від сукупної відповідності цілої ради дисертації. Виявлено, що значна частина членів рад мають не нульову схожість з дисертацією, 35 з них зовсім не відповідають тематиці дисертації для розподілу захищених дисертацій та 8 не відповідають тематиці для розподілу проблемних членів. Виявлено, що більша частина разових рад мають відповідність вище 0.2 (62 ради), і решта мають незначущий рівень відповідності (4 ради).

3. На основі отриманих даних проведено підбір опонентів до дисертацій за різних оптимізаційних підходів: повний перебір, повний перебір на обрізаній

множині кандидатів, жадібний алгоритм, жадібний алгоритм з елітизмом та ізольований підбір. За критерій обрано середнє покращення в порівнянні з разовою радою від закладу. Повний перебір має найкраще середнє покращення на рівні 55%, проте і час такого підбору найбільший. Виявлено, що деякі ради не були покращені або рівень покращення низький. Це зумовлено в першу чергу нерівномірністю розподілу освітніх спеціальностей дисертацій та малим розміром датасету. Встановлено, що перебір на обрізаній множині з рівнем 0.3 однозначно невдалий. Усі інші утворюють множину Парето. Тому для отримання оптимального результату необхідно враховувати баланс між цими величинами.

4. Подано детальний приклад підбору опонентів для однієї дисертації з датасету. В результаті підбору опонентів вдалося покращити відповідність разової ради на 46% в порівнянні з разовою радою від закладу. Встановлено, що індивідуальна схожість членів ради до дисертації може бути посередньою, проте сукупно рівень відповідності самої ради є вищим. Тобто, окремо кожен член ради не накриває всієї тематики дисертації, проте сукупно всі разом дають вищий рівень покриття.

Результати розділу опубліковано в роботах автора [19], [21].

ВИСНОВКИ

Усі поставлені завдання виконано. Мету дисертації досягнуто, тобто вирішено актуальну наукову задачу підвищення ефективності управління процесами експертизи наукових творів за рахунок розробки інформаційної технології підбору релевантних рецензентів. Наукову задачу вирішено шляхом розробки на базі принципів інтелектуального аналізу даних нового підходу до задачі підбору рецензентів та його реалізації у вигляді інформаційної технології. Розроблена інформаційна технологія являє собою взаємопов'язаний комплекс формалізованих постановок задач, критеріїв, моделей, алгоритмів і програм для підбору рецензентів на основі інтелектуального аналізу даних. Вона дозволяє ефективно вирішити задачу експрес-підбору колективу рецензентів з урахуванням якості рішення та обчислювальної складності перебору кандидатів. Розроблена інформаційна технологія ґрунтується на таких чотирьох наукових результатах.

1. Критерій оцінювання схожості потенційного рецензента та наукового твору.

Суть результату: удосконалено критерій Чекановського для оцінювання схожості категоріальних розподілів шляхом введення додаткового доданку, який враховує перехресну спорідненість категорій, що дозволяє взяти до уваги спорідненість наукових спеціальностей потенційного рецензента та наукового твору.

Новизна результату: вперше врахована перехресна схожість об'єктів, які описано розподілом на споріднених категоріях.

Достовірність результату підтверджена обчислювальними експериментами, які показали, що на тестовій вибірці доданок який враховує спорідненість категорій має значну кількість викидів як для малої так і для великої кількості категорій. Цей факт вказує на те, що запропонований критерій дозволяє легко ідентифікувати пари об'єктів, схожість яких визначається значною мірою через належність до споріднених категорій.

2. Модель статистичного оцінювання спорідненості наукових спеціальностей.

Суть результату: розроблено модель статистичного оцінювання спорідненості наукових спеціальностей, що ґрунтується на розрахунку індекса Жаккара за базою категоризованих наукових публікацій.

Новизна результату: вперше розроблено модель статистичного оцінювання спорідненості наукових спеціальностей, що дозволяє швидко ідентифікувати рівень схожості спеціальностей в порівнянні з трудомісткими аналогами на основі аналізу списку цитованості та списку джерел кожної окремої статті.

Достовірність результату підтверджена тестовими прикладами.

3. Метод категоризації рецензентів.

Суть результату: розроблено метод категоризації рецензентів, який базується на перетворенні обмеженого набору початкової інформації у формі наукових інтересів рецензента або ключових слів заявки у вектор в просторі наукових спеціальностей за принципом word2vec.

Новизна результату: вперше розроблено метод категоризації рецензентів, який відрізняється від формального співставлення шифрів спеціальностей аналізом фактичного змісту заявки та поточної компетентності рецензента, а від моделі співставлення ключових слів – врахуванням синонімії та спорідненості.

Достовірність результату забезпечена налаштуванням параметрів алгоритму категоризації на тестовому датасеті з 2489 науковців та експериментально підтверджена для спеціальностей з ANZSRC-2020 з використанням інформаційної бази системи Dimensions.

4. Метод призначення колективу рецензентів.

Суть результату: розроблено метод призначення колективу рецензентів.

Новизна результату: вперше розроблено метод призначення колективу рецензентів за їх відповідністю тематиці заявці, який, на відміну від ізолюваного підбору рецензентів враховує здатність саме колективу

рецензентів спільно оцінити заявку за усіма аспектами її тематики, та є збалансованим за критеріями якості підбору та витратами ресурсів на пошук колективу.

Достовірність результату підтверджена шляхом проведення експериментів з підбору рецензентів на тестовому датасеті з 67 PhD-дисертацій.

Практична цінність дисертації полягає в тому, що теоретичні складові інформаційної технології підбору рецензентів реалізовано програмно. Інформаційна технологія забезпечує ефективний експрес-підбір рецензентів з врахуванням здатності саме колективу рецензентів спільно оцінити заявку за усіма аспектами її тематики, та є збалансована за критеріями якості підбору і витратами ресурсів на пошук колективу. Практичну цінність підтверджує детальний приклад процесу підбору рецензентів в контексті задачі підбору опонентів до дисертацій. В результаті вдалося покращити склад разових рад, при порівнянні з радами від закладу, в середньому на 32-55% в залежності від типу алгоритму, що використовувався.

Розроблена інформаційна технологія може використовуватись для покращення ефективності управління процесів підбору рецензентів в таких сферах як: рецензування на конференціях; підбір рецензентів для експертизи статей поданих у журнали; підбір рецензентів для експертизи наукових проєктів поданих для отримання грантів; підбір рецензентів та опонентів для дисертацій; аудит разових рад дисертацій – перевірка коректності сформованої разової ради відповідно до тематики дисертації. При цьому необхідно врахувати, що розроблена інформаційна технологія не вирішує усі проблеми, які мають місце під час автоматичного підбору рецензентів, наприклад, у предметну область дослідження не входять питання усунення конфлікту інтересів, завантаженість рецензентів, формування актуального реєстру потенційних рецензентів тощо.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Датасет науковців для підбору параметрів алгоритму категоризації ключових слів. [Електронний ресурс]. Режим доступу: https://www.researchgate.net/publication/375744061_Researchers_dataset_for_categorization_Data_based_on_Dimensions_and_Google_Scholar.
2. Датасет PhD-дисертацій для проведення експериментів з підбору рецензентів за допомогою запропонованого методу підбору рецензентів. [Електронний ресурс]. Режим доступу: https://www.researchgate.net/publication/379726745_Dataset_for_PhD_theses_reviewers_assignments.
3. Shtovba, S., & Petrychko, M. (2019). Jaccard index-based assessing the similarity of research fields in dimensions. In CEUR Workshop Proceedings (Vol. 2533, pp. 117–128). CEUR-WS.
4. Shtovba, S., & Petrychko, M. (2021). An informetric assessment of various research fields interactions on base of categorized papers in dimensions. In CEUR Workshop Proceedings (Vol. 2845, pp. 159–169). CEUR-WS.
5. Petrychko M., Shtovba S., An Informetric Assessment of Various Research Fields Interactions on Base of Categorized Papers in Dimensions, in IT&I-2020: VII International conference «Information Technology and Interactions» – Satellite conference, Kyiv, 2020, pp. 170-171.
6. Петричко М. В., Штовба С. Д. Статистична оцінка рівня міждисциплінарності наукових досліджень з використанням даних системи Dimensions на «Матеріали наукової конференції професорсько-викладацького складу, наукових працівників і здобувачів наукового ступеня за підсумками науково-дослідної роботи за період 2019–2020 рр.», Донецький національний університет імені Василя Стуса, 2021.
7. Петричко М.В., Штовба С. Д. Метрика схожості категоріальних розподілів зі спорідненими категоріями на VIII міжнародній науково-технічній конференції «Комп’ютерне моделювання та оптимізації складних систем», Дніпро, 2023.

8. Петричко М. В., Штовба С. Д. Виявлення трендів мультидисциплінарних досліджень у 2022–2023 роках з використанням системи Dimensions, Прикладні аспекти сучасних міждисциплінарних досліджень, 200-202, Вінниця, 2024.

9. Штовба С. Д., Петричко М. В. (2024). Ідентифікація рівня спорідненості освітніх спеціальностей на основі аналізу профілей експертів НАЗЯВО, Наукові праці Вінницького національного технічного університету, №1. <https://doi.org/10.31649/2307-5376-2024-1-48-59>.

10. Штовба С. Д., Петричко М. В. (2024). Ідентифікація рівня спорідненості наукових спеціальностей на основі даних системи Dimensions, Проблеми програмування, вип. 1, 2024. <https://doi.org/10.15407/pp2024.01.077>.

11. Петричко М. В., Штовба С. Д. (2021). Модифікована метрика Чекановського для оцінки схожості науковців з врахуванням спорідненості спеціальностей, на Матеріали конференції «L Науково-технічна конференція підрозділів Вінницького національного технічного університету», Вінниця.

12. 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, pp. 299–311). CEUR-WS. <https://doi.org/10.32782/cmis/2864-26>.

13. Штовба С.Д., Петричко М.В. (2021). Тематичне моделювання науковців на основі їх інтересів у Google Scholar, Системні дослідження та інформаційні технології, №2, вересень, с. 113-129. <https://doi.org/10.20535/SRIT.2308-8893.2021.2.09>.

14. Петричко М. В., Штовба С. Д. Автоматична категоризація науковців за їх інтересами в Google Scholar, на XV міжнародній конференції «Контроль і управління в складних системах», Вінниця, 2020.

15. Петричко М. В., Штовба С. Д. Ідентифікація наукових спеціальностей дослідників на основі їх інтересів в Google Scholar, на «Матеріали II всеукраїнської науково-практичної конференції студентів,

аспірантів та молодих вчених «Прикладні інформаційні технології»: збірник наукових праць», Вінниця, Донецький національний університет імені Василя Стуса, 2021, с.70-73.

16. Петричко М. В., Штовба С. Д., Тилець Р. О. Автоматична категоризація науковців на основі профілей в Google Scholar з врахуванням спорідненості між науковими спеціальностями в Матеріали конференції «XLIX Науково-технічна конференція підрозділів Вінницького національного технічного університету (2020)», Вінниця, 2020.

17. Петричко М. В., Штовба С. Д., Биков М. М. Налаштування алгоритму категоризації дослідників з використанням інформаційних ресурсів системи Dimensions, на XVI міжнародній конференції «Контроль і управління в складних системах», Вінниця, 2022.

18. Штовба С.Д., Петричко М.В., Петранова М.Ю. (2023). Метрика схожості категоріальних розподілів, що враховує спорідненість різних категорій, Вісник Вінницького політехнічного інституту, №2, с. 49-57. <https://doi.org/10.31649/1997-9266-2023-167-2-49-57>.

19. 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, pp. 76–85). CEUR-WS. <https://doi.org/10.32782/cmis/3392-7>.

20. Петричко М. В., Штовба С. Д. (2024). Автоматизація підбору наукових рецензентів: огляд задач і методів, Вісник Вінницького політехнічного інституту, №1, с. 56–64. <https://doi.org/10.31649/1997-9266-2024-172-1-56-64>.

21. Штовба С.Д., Петричко М.В., Баєв А.В. Авторське свідоцтво на комп'ютерну програму «Double SimMetrics – метрика схожості категоріальних розподілів за подібних категорій», №116605 від 01.03.2023 р.

22. Shah, N. B. (2022). Challenges, experiments, and computational solutions in peer review. Communications of the ACM. Association for Computing Machinery. <https://doi.org/10.1145/3528086>.

23. Zhang, D., Zhao, S., Duan, Z., Chen, J., Zhang, Y., & Tang, J. (2019). A multi-label classification method using a hierarchical and transparent representation for paper-reviewer recommendation. arXiv. <https://doi.org/10.48550/arXiv.1912.08976>.
24. Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In proceedings of 20th conference on uncertainty in artificial intelligence (UAI), Banff, Canada pp. 487–494.
25. Dhull, K., Jecmen, S., Kothari, P., & Shah, N. B. (2022). Strategyproofing Peer Assessment via Partitioning: The Price in Terms of Evaluators' Expertise. arXiv preprint, arXiv:2201.10631. <https://doi.org/10.48550/arXiv.2201.10631>.
26. Karimzadehgan, M., & Zhai, C. (2012). Integer linear programming for Constrained Multi-Aspect Committee Review Assignment. Information Processing and Management, 48(4), 725–740. <https://doi.org/10.1016/j.ipm.2011.09.004>.
27. Концевой А. О., Бісікало О. В. Моделі глибокого навчання для вирішення задач класифікації текстової інформації / А. О. Концевой, О. В. Бісікало // Інформаційні технології та комп'ютерна інженерія. – 2022. – № 3. – С. 13–20. <https://doi.org/10.31649/1999-9941-2022-55-3-13-20>.
28. Xu, H., Kuchansky, A., Biloshchytska, S., & Tsiutsiura, M. (2021). A Conceptual Research Model for the Partner Selection Problem. In SIST 2021 - 2021 IEEE International Conference on Smart Information Systems and Technologies. Institute of Electrical and Electronics Engineers Inc. <https://doi.org/10.1109/SIST50301.2021.9465931>.
29. Price, S., & Flach, P. A. (2017). Computational support for academic peer review: A perspective from artificial intelligence. Communications of the ACM. Association for Computing Machinery. <https://doi.org/10.1145/2979672>.
30. Spier, R. (2002). The history of the peer-review process. Trends in Biotechnology. [https://doi.org/10.1016/S0167-7799\(02\)01985-6](https://doi.org/10.1016/S0167-7799(02)01985-6).
31. Brown, T. (2002). Peer Review and the Acceptance of New Scientific Ideas: Discussion Paper from a Working Party on Equipping the Public with an Understanding of Peer Review. Sense About Science.

32. Benos, D. J., Bashari, E., Chaves, J. M., Gaggar, A., Kapoor, N., LaFrance, M., Zotov, A. (2007). The ups and downs of peer review. *American Journal of Physiology - Advances in Physiology Education*. American Physiological Society. <https://doi.org/10.1152/advan.00104.2006>.
33. Jefferson, T., Alderson, P., Wager, E., & Davidoff, F. (2002). Effects of editorial peer review: A systematic review. *JAMA*. American Medical Association. <https://doi.org/10.1001/jama.287.21.2784>.
34. Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1). <https://doi.org/10.1057/s41599-021-00903-w>.
35. Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*. <https://doi.org/10.1002/asi.22784>.
36. Rennie, D. (2016). Let's make peer review scientific. *Nature* 535, 31–33. <https://doi.org/10.1038/535031a>.
37. Thorngate, W., & Chowdhury, W. (2014). By the numbers: Track record, flawed reviews, journal space, and the fate of talented authors. In *Advances in Intelligent Systems and Computing* (Vol. 229 AISC, pp. 177–188). Springer Verlag. https://doi.org/10.1007/978-3-642-39829-2_16.
38. Kelly J, Sadeghieh T, Adeli K. (2014). Peer Review in Scientific Publications: Benefits, Critiques, & A Survival Guide. *EJIFCC*. 25(3): 227-243.
39. 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>.
40. Ribeiro, A. C., Sizo, A., & Reis, L. P. (2023). Investigating the reviewer assignment problem: A systematic literature review. *Journal of Information Science*, 0(0). <https://doi.org/10.1177/01655515231176668>.

41. Kalmukov, Y. (2020). An algorithm for automatic assignment of reviewers to papers. *Scientometrics*, 124(3), 1811–1850. <https://doi.org/10.1007/s11192-020-03519-0>.
42. Pradhan, T., Sahoo, S., Singh, U., & Pal, S. (2021). A proactive decision support system for reviewer recommendation in academia. *Expert Systems with Applications*, 169. <https://doi.org/10.1016/j.eswa.2020.114331>.
43. Sun, Y. H., Ma, J., Fan, Z. P., & Wang, J. (2008). A hybrid knowledge and model approach for reviewer assignment. *Expert Systems with Applications*, 34(2), 817–824. <https://doi.org/10.1016/j.eswa.2006.10.021>.
44. Tayal, D. K., Saxena, P. C., Sharma, A., Khanna, G., & Gupta, S. (2014). New method for solving reviewer assignment problem using type-2 fuzzy sets and fuzzy functions. *Applied Intelligence*, 40(1), 54–73. <https://doi.org/10.1007/s10489-013-0445-5>.
45. Wang, F., Zhou, S., & Shi, N. (2013). Group-to-group reviewer assignment problem. *Computers and Operations Research*, 40(5), 1351–1362. <https://doi.org/10.1016/j.cor.2012.08.005>.
46. Goldsmith, J., & Sloan, R. H. (2007). The AI conference paper assignment problem. In *AAAI Workshop - Technical Report* (Vol. WS-07-10, pp. 53–57).
47. Rodriguez, M. A., & Bollen, J. (2008). An algorithm to determine peer-reviewers. In *International Conference on Information and Knowledge Management, Proceedings* (pp. 319–328). <https://doi.org/10.1145/1458082.1458127>.
48. Harper, P. R., De Senna, V., Vieira, I. T., & Shahani, A. K. (2005). A genetic algorithm for the project assignment problem. *Computers and Operations Research*, 32(5), 1255–1265. <https://doi.org/10.1016/j.cor.2003.11.003>.
49. Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology*, 45, 197–245. <https://doi.org/10.1002/aris.2011.1440450112>.
50. Fletcher, R.H., & Fletcher, S.W. (2003). 4 : The effectiveness of journal peer review.

51. Liu, O., Wang, J., Ma, J., & Sun, Y. (2016). An intelligent decision support approach for reviewer assignment in R&D project selection. *Computers in Industry*, 76, 1–10. <https://doi.org/10.1016/j.compind.2015.11.001>.
52. Protasiewicz, J. (2014). A support system for selection of reviewers. In *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics* (Vol. 2014-January, pp. 3062–3065). Institute of Electrical and Electronics Engineers Inc. <https://doi.org/10.1109/smc.2014.6974397>.
53. Kalmukov, Y. (2020). Automatic Assignment of Reviewers to Papers Based on Vector Space Text Analysis Model. In *ACM International Conference Proceeding Series* (pp. 229–235). Association for Computing Machinery. <https://doi.org/10.1145/3407982.3408026>.
54. Karimzadehgan, M., & Zhai, C. (2012). Integer linear programming for Constrained Multi-Aspect Committee Review Assignment. *Information Processing and Management*, 48(4), 725–740. <https://doi.org/10.1016/j.ipm.2011.09.004>.
55. Pradhan, D. K., Chakraborty, J., Choudhary, P., & Nandi, S. (2020). An automated conflict of interest based greedy approach for conference paper assignment system. *Journal of Informetrics*, 14(2). <https://doi.org/10.1016/j.joi.2020.101022>.
56. Yeşilçimen, A., & Yıldırım, E. A. (2019). An alternative polynomial-sized formulation and an optimization based heuristic for the reviewer assignment problem. *European Journal of Operational Research*, 276(2), 436–450. <https://doi.org/10.1016/j.ejor.2019.01.035>.
57. Dumais, S. T., & Nielsen, J. (1992). Automating the assignment of submitted manuscripts to reviewers. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 233–244). Publ by ACM. <https://doi.org/10.1145/133160.133205>.
58. 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>.

59. Wang, F., Chen, B., & Miao, Z. (2008). A survey on reviewer assignment problem. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5027 LNAI(60704048), 718–727. https://doi.org/10.1007/978-3-540-69052-8_75.

60. Xu, Y., Ma, J., Sun, Y., Hao, G., Xu, W., & Zhao, D. (2010). A decision support approach for assigning reviewers to proposals. *Expert Systems with Applications*, 37(10), 6948–6956. <https://doi.org/10.1016/j.eswa.2010.03.027>.

61. Kou, N. M., Hou, L. U., Mamoulis, N., & Gong, Z. (2015). Weighted coverage based reviewer assignment. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (Vol. 2015-May, pp. 2031–2046)*. Association for Computing Machinery. <https://doi.org/10.1145/2723372.2723727>.

62. Duan, Z., Tan, S., Zhao, S., Wang, Q., Chen, J., & Zhang, Y. (2019). Reviewer assignment based on sentence pair modeling. *Neurocomputing*, 366, 97–108. <https://doi.org/10.1016/j.neucom.2019.06.074>.

63. Yang, C., Liu, T., Yi, W., Chen, X., & Niu, B. (2020). Identifying expertise through semantic modeling: A modified BBPSO algorithm for the reviewer assignment problem. *Applied Soft Computing Journal*, 94, 106483. <https://doi.org/10.1016/j.asoc.2020.106483>.

64. Mimno, D., & McCallum, A. (2007). Expertise modeling for matching papers with reviewers. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 500–509)*. <https://doi.org/10.1145/1281192.1281247>.

65. Benferhat, S., & Lang, J. (2001). Conference paper assignment. *International Journal of Intelligent Systems*, 16(10), 1183–1192. <https://doi.org/10.1002/int.1055>.

66. Cook, W. D., Golany, B., Kress, M., Penn, M., & Raviv, T. (2005). Optimal allocation of proposals to reviewers to facilitate effective ranking. *Management Science*, 51(4), 655–661. <https://doi.org/10.1287/mnsc.1040.0290>.

67. Kolasa, T., & Krol, D. (2011). A survey of algorithms for paper-reviewer assignment problem. *IETE Technical Review (Institution of Electronics and*

Telecommunication Engineers, India), 28(2), 123–134. <https://doi.org/10.4103/0256-4602.78092>.

68. 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).

69. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference (Vol. 1, pp. 2227–2237). Association for Computational Linguistics (ACL). <https://doi.org/10.18653/v1/n18-1202>.

70. Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever P. Language Models are Unsupervised Multitask Learners / A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, P. Sutskever / Radford. – 2018.

71. Howard J., Ruder S. Universal Language Model Fine-tuning for Text Classification / J. Howard, S. Ruder // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. – 2018.

72. 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>.

73. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems. Neural information processing systems foundation.

74. Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural Probabilistic Language Model. In *Journal of Machine Learning Research* (Vol. 3, pp. 1137–1155). <https://doi.org/10.1162/153244303322533223>.
75. Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5), 513–523. [https://doi.org/10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0).
76. Arora, S., Ge, R., & Moitra, A. (2012). Learning topic models - Going beyond SVD. In *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS* (pp. 1–10). <https://doi.org/10.1109/FOCS.2012.49>.
77. 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>.
78. Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999* (pp. 50–57). Association for Computing Machinery, Inc. <https://doi.org/10.1145/312624.312649>.
79. 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.
80. Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*. <https://doi.org/10.1108/eb026526>.
81. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. International Conference on Learning Representations, ICLR.
82. Arora, S., Ge, R., & Moitra, A. (2012). Learning topic models - Going beyond SVD. In *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS* (pp. 1–10). <https://doi.org/10.1109/FOCS.2012.49>.

83. Benferhat, S., & Lang, J. (2001). Conference paper assignment. *International Journal of Intelligent Systems*, 16(10), 1183–1192. <https://doi.org/10.1002/int.1055>.

84. Basu, C., Hirsh, H., Cohen, W. W., & Nevill-Manning, C. (2001). Technical paper recommendation: A study in combining multiple information sources. *Journal of Artificial Intelligence Research*, 14, 241–262. <https://doi.org/10.1613/jair.739>.

85. Janak, S. L., Taylor, M. S., Floudas, C. A., Burka, M., & Mountziaris, T. J. (2006). Novel and effective integer optimization approach for the NSF panel-assignment problem: A multiresource and preference-constrained generalized assignment problem. *Industrial and Engineering Chemistry Research*, 45(1), 258–265. <https://doi.org/10.1021/ie050478h>.

86. Papagelis, M., Plexousakis, D., & Nikolaou, P. N. (2005). CONFIOUS: Managing the electronic submission and reviewing process of scientific conferences. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 3806 LNCS, pp. 711–720). https://doi.org/10.1007/11581062_78.

87. Garg, N., Kavitha, T., Kumar, A., Mehlhorn, K., & Mestre, J. (2010). Assigning papers to referees. *Algorithmica* (New York), 58(1), 119–136. <https://doi.org/10.1007/s00453-009-9386-0>.

88. Xu, Y. H., Guo, X. T., Xu, L., Chen, Y., & Zhuang, Y. Y. (2012). Research analytics for reviewer recommendation. In *International Conference on Management Science and Engineering - Annual Conference Proceedings* (pp. 213–217). <https://doi.org/10.1109/ICMSE.2012.6414185>.

89. Daş, G. S., & Göçken, T. (2014). A fuzzy approach for the reviewer assignment problem. *Computers and Industrial Engineering*, 72(1), 50–57. <https://doi.org/10.1016/j.cie.2014.02.014>.

90. Liu, D., Xu, W., Du, W., & Wang, F. (2015). How to choose appropriate experts for peer review: An intelligent recommendation method in a big data context.

In Data Science Journal (Vol. 14). Committee on Data for Science and Technology.
<https://doi.org/10.5334/dsj-2015-016>.

91. Jecmen, S., Zhang, H., Liu, R., Shah, N. B., Conitzer, V., & Fang, F. (2020). Mitigating Manipulation in Peer Review via Randomized Reviewer Assignments. arXiv preprint arXiv:2006.16437. <http://arxiv.org/abs/2006.16437>.

92. Fiez, T., Shah, N. B., & Ratliff, L. (2020). A SUPER* Algorithm to Optimize Paper Bidding in Peer Review. In Proceedings of Machine Learning Research (Vol. 124, pp. 580–589). ML Research Press..

93. Cabanac, G., & Preuss, T. (2013). Capitalizing on order effects in the bids of peer-reviewed conferences to secure reviews by expert referees. In Journal of the American Society for Information Science and Technology (Vol. 64, pp. 405–415).
<https://doi.org/10.1002/asi.22747>.

94. Guo, L., Wu, J., Chang, W., Wu, J., & Li, J. (2018). K-Loop Free Assignment in Conference Review Systems. In 2018 International Conference on Computing, Networking and Communications, ICNC 2018 (pp. 542-547). (2018 International Conference on Computing, Networking and Communications, ICNC 2018). Institute of Electrical and Electronics Engineers Inc..
<https://doi.org/10.1109/ICCNC.2018.8390343>.

95. Hartvigsen, D., Wei, J. C., & Czuchlewski, R. (1999). The conference paper-reviewer assignment problem. In Decision Sciences (Vol. 30, pp. 865–876). Decision Sciences Institute. <https://doi.org/10.1111/j.1540-5915.1999.tb00910.x>.

96. Charlin, L., Zemel, R., & Boutilier, C. (2012). A framework for optimizing paper matching. In proceedings of the 27th Conference on Uncertainty in Artificial Intelligence, 86–95.

97. Liu, X., Suel, T., & Memon, N. (2014). A Robust Model for Paper-Reviewer Assignment Categories and Subject Descriptors. In RecSys 2014 - Proceedings of the 8th ACM Conference on Recommender Systems2014 (pp. 25–32).

98. Silva, T., Jian, M., & Chen, Y. (2014). Process analytics approach for R&D project selection. *ACM Transactions on Management Information Systems*, 5(4). <https://doi.org/10.1145/2629436>.
99. Davoodi, E., Kianmehr, K., & Afsharchi, M. (2013). A semantic social network-based expert recommender system. *Applied Intelligence*, 39(1), 1–13. <https://doi.org/10.1007/s10489-012-0389-1>.
100. 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>.
101. 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 NLP and Very-Large Corpora*.
102. Hettich, S., & Pazzani, M. J. (2006). Mining for proposal reviewers: Lessons learned at the national science foundation. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Vol. 2006, pp. 862–871)*.
103. Caldera, C., Eggeling, E., Schr, M., & Fellner, D. W.. Towards an Automatic Review/Paper Matching degree Calculation. In *Proceedings of CONTENT 2014: The Sixth International Conference on Creative Content Technologies*, 70–75, 2014.
104. Biswas, H. K., & Hasan, M. M. (2007). Using publications and domain knowledge to build research profiles: An application in automatic reviewer assignment. In *ICICT 2007: Proceedings of International Conference on Information and Communication Technology* (pp. 82–86). <https://doi.org/10.1109/ICICT.2007.375347>.
105. Shon, H. S., Han, S. H., Kim, K. A., Cha, E. J., & Ryu, K. H. (2017, April 1). Proposal reviewer recommendation system based on big data for a national research management institute. *Journal of Information Science*. SAGE Publications Ltd. <https://doi.org/10.1177/0165551516644168>.

106. Mai, X., Ding, G., & Wang, J. (2012). Authority aware expert search: Algorithm and system for NSFC. In IET Conference Publications (Vol. 2012, pp. 585–588). <https://doi.org/10.1049/cp.2012.1047>.
107. Charlin, L., & Zemel, R. S. (2013). The Toronto Paper Matching System: An automated paper-reviewer assignment system. In ICML Workshop on Peer Reviewing and Publishing Models (PEER) (Vol. 28).
108. Karimzadehgan, M., Zhai, C. X., & Belford, G. (2008). Multi-aspect expertise matching for review assignment. In International Conference on Information and Knowledge Management, Proceedings (pp. 1113–1122). <https://doi.org/10.1145/1458082.1458230>.
109. 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>.
110. Karimzadehgan, M., & Zhai, C. X. (2009). Constrained multi-aspect expertise matching for committee review assignment. In International Conference on Information and Knowledge Management, Proceedings (pp. 1697–1700). <https://doi.org/10.1145/1645953.1646207>.
111. Tang, W., Tang, J., Lei, T., Tan, C., Gao, B., & Li, T. (2012). On optimization of expertise matching with various constraints. Neurocomputing, 76(1), 71–83. <https://doi.org/10.1016/j.neucom.2011.04.039>.
112. Kou, N. M., Leong Hou, U., Mamoulis, N., Li, Y., Li, Y., & Gong, Z. (2015). A topic-based reviewer assignment system. In Proceedings of the VLDB Endowment (Vol. 8, pp. 1852–1855). Association for Computing Machinery. <https://doi.org/10.14778/2824032.2824084>.
113. Jin, J., Geng, Q., Mou, H., & Chen, C. (2019). Author–Subject–Topic model for reviewer recommendation. Journal of Information Science, 45(4), 554–570. <https://doi.org/10.1177/0165551518806116>.
114. Kawamae, N. (2010). Author interest topic model. In SIGIR 2010 Proceedings - 33rd Annual International ACM SIGIR Conference on Research and

Development in Information Retrieval (pp. 887–888).
<https://doi.org/10.1145/1835449.1835666>.

115. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407. [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6<391::AID-ASI1>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9).

116. Li, B., & Hou, Y. T. (2016, September 1). The new automated IEEE INFOCOM review assignment system. *IEEE Network*. Institute of Electrical and Electronics Engineers Inc. <https://doi.org/10.1109/MNET.2016.7579022>.

117. Farahat, A., & Chen, F. (2006). Improving probabilistic latent semantic analysis with principal component analysis. In *EACL 2006 - 11th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference* (pp. 105–112)..

118. Tang, W., Tang, J., & Tan, C. (2010). Expertise matching via constraint-based optimization. In *Proceedings - 2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010* (Vol. 1, pp. 34–41). <https://doi.org/10.1109/WI-IAT.2010.133>.

119. Yunhong, X., & Xianli, Z. (2016). A LDA model based text-mining method to recommend reviewer for proposal of research project selection. In *2016 13th International Conference on Service Systems and Service Management, ICSSSM 2016*. Institute of Electrical and Electronics Engineers Inc. <https://doi.org/10.1109/ICSSSM.2016.7538568>.

120. Peng, H., Hu, H., Wang, K., & Wang, X. (2017). Time-aware and topic-based reviewer assignment. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 10179 LNCS, pp. 145–157). Springer Verlag. https://doi.org/10.1007/978-3-319-55705-2_11.

121. Yan, S. (2017). Utilizing Academic-Network-Based Conflict of Interests for Paper Reviewer Assignment. *International Journal of Knowledge Engineering*, 65–73. <https://doi.org/10.18178/ijke.2017.3.2.089>.

122. Nguyen, J., Sánchez-Hernández, G., Agell, N., Rovira, X., & Angulo, C. (2018). A decision support tool using Order Weighted Averaging for conference review assignment. *Pattern Recognition Letters*, 105, 114–120. <https://doi.org/10.1016/j.patrec.2017.09.020>.

123. Medakene, A. N., Bouanane, K., & Eddoud, M. A. (2019). A New Approach for Computing the Matching Degree in the Paper-to-Reviewer Assignment Problem. In 2019 International Conference on Theoretical and Applicative Aspects of Computer Science, ICTAACS 2019. Institute of Electrical and Electronics Engineers Inc. <https://doi.org/10.1109/ICTAACS48474.2019.8988127>.

124. Parikshit N. Mahalle, A. H. P. (2021). Fair and Accurate Review in Publication Process: A learning-based Proactive Approach for Assigning Reviewers to Manuscripts. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(4), 7–15. <https://doi.org/10.17762/turcomat.v12i4.455>.

125. Hoang, D. T., Nguyen, N. T., Collins, B., & Hwang, D. (2021). Decision Support System for Solving Reviewer Assignment Problem. *Cybernetics and Systems*, 52(5), 379–397. <https://doi.org/10.1080/01969722.2020.1871227>.

126. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner: Extraction and mining of academic social networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 990–998). <https://doi.org/10.1145/1401890.1402008>.

127. Daud, A., Li, J., Zhou, L., & Muhammad, F. (2010). Temporal expert finding through generalized time topic modeling. *Knowledge-Based Systems*, 23(6), 615–625. <https://doi.org/10.1016/j.knosys.2010.04.008>.

128. Ekinici, 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>.

129. 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>.

130. 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.

131. 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>.

132. 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).

133. 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.

134. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

135. 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>.

136. Zhao Y., Tang J., Du Z. (2019) EFCNN: A Restricted Convolutional Neural Network for Expert Finding. In: Yang Q., Zhou ZH., Gong Z., Zhang ML., Huang SJ. (eds) *Advances in Knowledge Discovery and Data Mining*. PAKDD 2019.

Lecture Notes in Computer Science, vol 11440. Springer, Cham.
https://doi.org/10.1007/978-3-030-16145-3_8.

137. Zhang, D., Zhao, S., Duan, Z., Chen, J., Zhang, Y., & Tang, J. (2019). A multi-label classification method using a hierarchical and transparent representation for paper-reviewer recommendation. arXiv.
<https://doi.org/10.48550/arXiv.1912.08976>.

138. Long, C., Wong, R. C. W., Peng, Y., & Ye, L. (2013). On good and fair paper-reviewer assignment. In proceedings IEEE International Conference on Data Mining, 1145–1150. <https://doi.org/10.1109/ICDM.2013.13>.

139. Taylor, C. J. (2008). On the optimal assignment of conference papers to reviewers. Technical Report MS-CIS-08-30, University of Pennsylvania.

140. Xue, N., Hao, J. X., Jia, S. L., & Wang, Q. (2012). An interval fuzzy ontology-based peer review assignment method. In proceedings of IEEE International Conference on E-Business Engineering, 55–60.
<https://doi.org/10.1109/ICEBE.2012.19>.

141. Kale, A., Kharat, R., Bodkhe, S., Apte, P., & Dhonde, H. (2015). Automated fair paper reviewer assignment for conference management system. In Proceedings of 1st International Conference on Computing, Communication, Control and Automation, ICCUBEA 2015, 408–411.
<https://doi.org/10.1109/ICCUBEA.2015.85>.

142. Stelmakh, I., Shah, N. B., & Singh, A. (2019). PeerReview4All: Fair and Accurate Reviewer Assignment in Peer Review. In Proceedings of Machine Learning Research (Vol. 98, pp. 828–856). ML Research Press.

143. Kobren, A., Saha, B., & McCallum, A. (2019). Paper matching with local fairness constraints. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1247–1257). Association for Computing Machinery. <https://doi.org/10.1145/3292500.3330899>.

144. Payan, J., & Zick, Y. (2022). I Will Have Order! Optimizing Orders for Fair Reviewer Assignment. In IJCAI International Joint Conference on Artificial

Intelligence (pp. 440–446). International Joint Conferences on Artificial Intelligence. <https://doi.org/10.24963/ijcai.2022/63>.

145. Alon, N., Fischer, F., Procaccia, A., & Tennenholtz, M. (2011). Sum of us: Strategyproof selection from the selectors. In ACM International Conference Proceeding Series (pp. 101–110). <https://doi.org/10.1145/2000378.2000390>.

146. Boehmer, N., Bredereck, R., & Nichterlein, A. (2022). Combating Collusion Rings is Hard but Possible. In Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022 (Vol. 36, pp. 4843–4850). Association for the Advancement of Artificial Intelligence. <https://doi.org/10.1609/aaai.v36i5.20412>.

147. Leyton-Brown, K., Mausam, Nandwani, Y., Zarkoob, H., Cameron, C., Newman, N., Raghu, D. (2022). Matching Papers and Reviewers at Large Conferences. Artificial Intelligence, Volume 331, ISSN 0004-3702. <https://doi.org/10.1016/j.artint.2024.104119>.

148. Merelo-Guervós, J. J., & Castillo-Valdivieso, P. (2004). Conference paper assignment using a combined greedy/evolutionary algorithm. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 3242, 602–611. https://doi.org/10.1007/978-3-540-30217-9_61.

149. Janak, S. L., Taylor, M. S., Floudas, C. A., Burka, M. K., & Mountziaris, T. J. (2005). A novel and effective integer optimization approach for the Nsf panel assignment problem: A multi-resource and preference-constrained generalized assignment problem. In Proceedings of AIChE Annual Meeting, 609, 6066.

150. Li, N. N., Zhao, Z., Gu, J. H., & Liu, B. Y. (2008). Ant colony optimization algorithm for expert assignment problem. In proceedings of the 7th International Conference on Machine Learning and Cybernetics, 2(July), 660–664. <https://doi.org/10.1109/ICMLC.2008.4620487>.

151. Conry, D., Koren, Y., & Ramakrishnan, N. (2009). Recommender systems for the conference paper assignment problem. In proceedings of the 3rd ACM Conference on Recommender Systems, 357–360. <https://doi.org/10.1145/1639714.1639787>.

152. Kat, B. (2021). An algorithm and a decision support system for the panelist assignment problem: The case of TUBITAK. *Journal of the Faculty of Engineering and Architecture of Gazi University*, 36(1), 69–87. <https://doi.org/10.17341/gazimmfd.631071>.

153. Li, N. N., Zhang, J. N., Gu, J. H., & Liu, B. Y. (2007). Solving expert assignment problem using improved genetic algorithm. In *proceedings of the Sixth International Conference on Machine Learning and Cybernetics*, 2(August), 934–937. <https://doi.org/10.1109/ICMLC.2007.4370276>.

154. Li, J., Peng, J., & Wei, Y. (2013). Adaptive parallel genetic algorithm for expert assignment problem. In *Proceedings of 6th International Symposium on Computational Intelligence and Design*, 1, 23–26. <https://doi.org/10.1109/ISCID.2013.13>.

155. Huang, M., Liu, B., & Hong, L. (2010). On assigning papers to reviewers. *2nd International Workshop on Database Technology and Applications*. <https://doi.org/10.1109/DBTA.2010.5659009>.

156. Штовба С.Д., Мазуренко В.В. Інтелектуальні технології ідентифікації залежностей. *Лабораторний практикум: електронний навчальний посібник*. – Вінниця: Вінницький національний технічний університет, 2014. – 113 с.

157. Штовба С.Д., Козачко О.М., *Machine learning: стартовий курс : електронний навчальний посібник* / Штовба С.Д., Козачко О.М. – Вінниця : ВНТУ, 2020. – 81 с. <https://iq.vntu.edu.ua/repository/card.php?id=1404>.

158. 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>.

159. Wang, W. J. (1997). New similarity measures on fuzzy sets and on elements. *Fuzzy Sets and Systems*, 85(3), 305–309. [https://doi.org/10.1016/0165-0114\(95\)00365-7](https://doi.org/10.1016/0165-0114(95)00365-7).

160. Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1(2), 1.
161. Jie, Y., Amores, J., Sebe, N., & Qi, T. (2006). Toward robust distance metric analysis for similarity estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 1, pp. 316–322). <https://doi.org/10.1109/CVPR.2006.310>.
162. Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., Rafols I., Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics*, 5(1), 14–26. <https://doi.org/10.1016/j.joi.2010.06.004>.
163. Dias, L., Gerlach, M., Scharloth, J., & Altmann, E. G. (2018). Using text analysis to quantify the similarity and evolution of scientific disciplines. *Royal Society Open Science*, 5(1). <https://doi.org/10.1098/rsos.171545>.
164. Porter, A. L., Cohen, A. S., David Roessner, J., & Perreault, M. (2007). Measuring researcher interdisciplinarity. *Scientometrics*, 72(1), 117–147. <https://doi.org/10.1007/s11192-007-1700-5>.
165. Van Noorden, R. (2015). Interdisciplinary research by the numbers. *Nature*. Nature Publishing Group. <https://doi.org/10.1038/525306a>.
166. Karlovčec, M., & Mladenović, D. (2015). Interdisciplinarity of scientific fields and its evolution based on graph of project collaboration and co-authoring. *Scientometrics*, 102(1), 433–454. <https://doi.org/10.1007/s11192-014-1355-y>.
167. Braam, R. R., Moed, H. F., & van Raan, A. F. J. (1991). Mapping of science by combined co-citation and word analysis. II: Dynamical aspects. *Journal of the American Society for Information Science*, 42(4), 252–266. [https://doi.org/10.1002/\(SICI\)1097-4571\(199105\)42:4<252::AID-ASI2>3.0.CO;2-G](https://doi.org/10.1002/(SICI)1097-4571(199105)42:4<252::AID-ASI2>3.0.CO;2-G).
168. Silva, F. N., Amancio, D. R., Bardosova, M., Costa, L. da F., & Oliveira, O. N. (2016). Using network science and text analytics to produce surveys in a scientific topic. *Journal of Informetrics*, 10(2), 487–502. <https://doi.org/10.1016/j.joi.2016.03.008>.

169. Dimensions API request. Електронний ресурс. Режим доступу: <https://www.dimensions.ai/products/all-products/dimensions-api/> (20.02.2024).

170. Dimensions DSL. Електронний ресурс. Режим доступу: <https://docs.dimensions.ai/dsl/> (20.02.2024).

171. Carbonneau, M. A., Cheplygina, V., Granger, E., & Gagnon, G. (2018). Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77, 329–353. <https://doi.org/10.1016/j.patcog.2017.10.009>.

172. НАЗЯВО, Захисти дисертацій. Електронний ресурс. Режим доступу <https://svr.naqa.gov.ua/#/all-defenses> (20.02.2024).

ДОДАТКИ

ДОДАТОК А. ПЕРЕЛІК ПУБЛІКАЦІЙ ЗДОБУВАЧА

- [1] Shtovba, S., & Petrychko, M. (2019). Jaccard index-based assessing the similarity of research fields in dimensions. In CEUR Workshop Proceedings (Vol. 2533, pp. 117–128). CEUR-WS. (Входить до Scopus)
- [2] Shtovba, S., & Petrychko, M. (2021). An informetric assessment of various research fields interactions on base of categorized papers in dimensions. In CEUR Workshop Proceedings (Vol. 2845, pp. 159–169). CEUR-WS. (Входить до Scopus)
- [3] 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, pp. 299–311). CEUR-WS. <https://doi.org/10.32782/cmis/2864-26>. (Входить до Scopus)
- [4] Штовба С.Д., Петричко М.В. (2021). Тематичне моделювання науковців на основі їх інтересів у Google Scholar, Системні дослідження та інформаційні технології, №2, вересень, с. 113-129. ISSN: 1681-6048. <https://doi.org/10.20535/SRIT.2308-8893.2021.2.09>. (Входить до Scopus)
- [5] Штовба С.Д., Петричко М.В., Петранова М.Ю. (2023). Метрика схожості категоріальних розподілів, що враховує спорідненість різних категорій, Вісник Вінницького політехнічного інституту, №2, с. 49-57. <https://doi.org/10.31649/1997-9266-2023-167-2-49-57>. (Фахове видання)
- [6] 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, pp. 76–85). CEUR-WS. <https://doi.org/10.32782/cmis/3392-7>. (Входить до Scopus).
- [7] Штовба С. Д., Петричко М. В. (2024). Ідентифікація рівня спорідненості освітніх спеціальностей на основі аналізу профілей експертів НАЗЯВО. Наукові праці Вінницького національного технічного університету, №1. <https://doi.org/10.31649/2307-5376-2024-1-48-59>. (Фахове видання)
- [8] Штовба С. Д., Петричко М. В. (2024). Ідентифікація рівня спорідненості наукових спеціальностей на основі даних системи Dimensions.

Проблеми програмування, №1. <https://doi.org/10.15407/pp2024.01.077>. (Фахове видання)

[9] Петричко М. В., Штовба С. Д. (2024). Автоматизація підбору наукових рецензентів: огляд задач і методів. Вісник Вінницького політехнічного інституту, №1, с. 56–64. <https://doi.org/10.31649/1997-9266-2024-172-1-56-64>. (Фахове видання)

[10] Петричко М. В., Штовба С. Д. (2020). Автоматична категоризація науковців за їх інтересами в Google Scholar, на XV міжнародній конференції «Контроль і управління в складних системах», Вінниця.

[11] Petrychko, M., Shtovba S. (2020). An Informetric Assessment of Various Research Fields Interactions on Base of Categorized Papers in Dimensions, in IT&I-2020: VII International conference «Information Technology and Interactions» – Satellite conference, Kyiv, pp. 170-171.

[12] Петричко М. В., Штовба С. Д. (2021). Статистична оцінка рівня міждисциплінарності наукових досліджень з використанням даних системи Dimensions, на «Матеріали наукової конференції професорсько-викладацького складу, наукових працівників і здобувачів наукового ступеня за підсумками науково-дослідної роботи за період 2019–2020 рр.», Донецький національний університет імені Василя Стуса.

[13] Петричко М. В., Штовба С. Д. (2021). Ідентифікація наукових спеціальностей дослідників на основі їх інтересів в Google Scholar, на Матеріали II всеукраїнської науково-практичної конференції студентів, аспірантів та молодих вчених «Прикладні інформаційні технології», Вінниця, Донецький національний університет імені Василя Стуса, с.70-73.

[14] Петричко М. В., Штовба С. Д., Биков М.М. (2022). Налаштування алгоритму категоризації дослідників з використанням інформаційних ресурсів системи Dimensions, на XVI міжнародній конференції «Контроль і управління в складних системах», Вінниця.

[15] Петричко М. В., Штовба С. Д. (2023). Метрика схожості категоріальних розподілів зі спорідненими категоріями, на VIII міжнародній

науково-технічній конференції «Комп'ютерне моделювання та оптимізації складних систем», Дніпро.

[16] Петричко М. В., Штовба С. Д. (2024). Виявлення трендів мультидисциплінарних досліджень у 2022–2023 роках з використанням системи Dimensions. Прикладні аспекти сучасних міждисциплінарних досліджень, 200–202, Вінниця.

[17] Петричко М. В., Штовба С. Д. (2021). Модифікована метрика Чекановського для оцінки схожості науковців з врахуванням спорідненості спеціальностей, на Матеріали конференції «L Науково-технічна конференція підрозділів Вінницького національного технічного університету», Вінниця.

[18] Петричко М. В., Штовба С. Д., Тилець Р. О. (2020). Автоматична категоризація науковців на основі профілей в Google Scholar з врахуванням спорідненості між науковими спеціальностями, на Матеріали конференції «XLIX Науково-технічна конференція підрозділів Вінницького національного технічного університету», Вінниця.

[19] Штовба С.Д., Петричко М.В., Баєв А.В. Авторське свідоцтво на комп'ютерну програму «Double SimMetrics – метрика схожості категоріальних розподілів за подібних категорій», №116605 від 01.03.2023 р.

ДОДАТОК Б. РЕЗУЛЬТАТИ ІДЕНТИФІКАЦІЇ СПОРІДНЕНOSTI
СПЕЦІАЛЬНОСТЕЙ З ANZSRC-2020

- 30 Agricultural, Veterinary and Food Sciences;
 - 3001 Agricultural Biotechnology;
 - 3002 Agriculture, Land and Farm Management;
 - 3003 Animal Production;
 - 3004 Crop and Pasture Production;
 - 3005 Fisheries Sciences;
 - 3006 Food Sciences;
 - 3007 Forestry Sciences;
 - 3008 Horticultural Production;
 - 3009 Veterinary Sciences;
- 31 Biological Sciences;
 - 3101 Biochemistry and Cell Biology;
 - 3102 Bioinformatics and Computational Biology;
 - 3103 Ecology;
 - 3104 Evolutionary Biology;
 - 3105 Genetics;
 - 3106 Industrial Biotechnology;
 - 3107 Microbiology;
 - 3108 Plant Biology;
 - 3109 Zoology;
- 32 Biomedical and Clinical Sciences;
 - 3201 Cardiovascular Medicine and Haema-tology;
 - 3202 Clinical Sciences;
 - 3203 Dentistry;
 - 3204 Immunology;
 - 3205 Medical Biochemistry and Metabolomics;
 - 3206 Medical Biotechnology;
 - 3207 Medical Microbiology;

- 3208 Medical Physiology;
- 3209 Neurosciences;
- 3210 Nutrition and Dietetics;
- 3211 Oncology and Carcinogenesis;
- 3212 Ophthalmology and Optometry;
- 3213 Paediatrics;
- 3214 Pharmacology and Pharmaceutical Sciences;
- 3215 Reproductive Medicine;
- 33 Built Environment and Design;
- 3301 Architecture;
- 3302 Building;
- 3303 Design;
- 3304 Urban and Regional Planning;
- 34 Chemical Sciences;
- 3401 Analytical Chemistry;
- 3402 Inorganic Chemistry;
- 3403 Macromolecular and Materials Chemistry;
- 3404 Medicinal and Biomolecular Chemistry;
- 3405 Organic Chemistry;
- 3406 Physical Chemistry;
- 3407 Theoretical and Computational Chemistry;
- 35 Commerce, Management, Tourism and Services;
- 3501 Accounting, Auditing and Accountability;
- 3502 Banking, Finance and Investment;
- 3503 Business Systems in Context;
- 3504 Commercial Services;
- 3505 Human Resources and Industrial Relations;
- 3506 Marketing;
- 3507 Strategy, Management and Organisational Behaviour;
- 3508 Tourism;

3509 Transportation, Logistics and Supply Chains;
36 Creative Arts and Writing;
3601 Art History, Theory and Criticism;
3602 Creative and Professional Writing;
3603 Music;
3604 Performing Arts;
3605 Screen and Digital Media;
3606 Visual Arts;
37 Earth Sciences;
3701 Atmospheric Sciences;
3702 Climate Change Science;
3703 Geochemistry;
3704 Geoinformatics;
3705 Geology;
3706 Geophysics;
3707 Hydrology;
3708 Oceanography;
3709 Physical Geography and Environmental Geoscience;
38 Economics;
3801 Applied Economics;
3802 Econometrics;
3803 Economic Theory;
39 Education;
3901 Curriculum and Pedagogy;
3902 Education Policy, Sociology and Philosophy;
3903 Education Systems;
3904 Specialist Studies in Education;
40 Engineering;
4001 Aerospace Engineering;
4002 Automotive Engineering;

- 4003 Biomedical Engineering;
- 4004 Chemical Engineering;
- 4005 Civil Engineering;
- 4006 Communications Engineering;
- 4007 Control Engineering, Mechatronics and Robotics;
- 4008 Electrical Engineering;
- 4009 Electronics, Sensors and Digital Hardware;
- 4010 Engineering Practice and Education;
- 4011 Environmental Engineering;
- 4012 Fluid Mechanics and Thermal Engineering;
- 4013 Geomatic Engineering;
- 4014 Manufacturing Engineering;
- 4015 Maritime Engineering;
- 4016 Materials Engineering;
- 4017 Mechanical Engineering;
- 4018 Nanotechnology;
- 4019 Resources Engineering and Extractive Metallurgy;
- 41 Environmental Sciences;
- 4101 Climate Change Impacts and Adaptation;
- 4102 Ecological Applications;
- 4103 Environmental Biotechnology;
- 4104 Environmental Management;
- 4105 Pollution and Contamination;
- 4106 Soil Sciences;
- 42 Health Sciences;
- 4201 Allied Health and Rehabilitation Science;
- 4202 Epidemiology;
- 4203 Health Services and Systems;
- 4204 Midwifery;
- 4205 Nursing;

4206 Public Health;
4207 Sports Science and Exercise;
4208 Traditional, Complementary and Integrative Medicine;
43 History, Heritage and Archaeology;
4301 Archaeology;
4302 Heritage, Archive and Museum Studies;
4303 Historical Studies;
44 Human Society;
4401 Anthropology;
4402 Criminology;
4403 Demography;
4404 Development Studies;
4405 Gender Studies;
4406 Human Geography;
4407 Policy and Administration;
4408 Political Science;
4409 Social Work;
4410 Sociology;
46 Information and Computing Sciences;
4601 Applied Computing;
4602 Artificial Intelligence;
4603 Computer Vision and Multimedia Computation;
4604 Cybersecurity and Privacy;
4605 Data Management and Data Science;
4606 Distributed Computing and Systems Software;
4607 Graphics, Augmented Reality and Games;
4608 Human-Centred Computing;
4609 Information Systems;
4610 Library and Information Studies;
4611 Machine Learning;

4612 Software Engineering;
4613 Theory of Computation;
47 Language, Communication and Culture;
4701 Communication and Media Studies;
4702 Cultural Studies;
4703 Language Studies;
4704 Linguistics;
4705 Literary Studies;
48 Law and Legal Studies;
4801 Commercial Law;
4802 Environmental and Resources Law;
4803 International and Comparative Law;
4804 Law in Context;
4805 Legal Systems;
4806 Private Law and Civil Obligations;
4807 Public Law;
49 Mathematical Sciences;
4901 Applied Mathematics;
4902 Mathematical Physics;
4903 Numerical and Computational Mathematics;
4904 Pure Mathematics;
4905 Statistics;
50 Philosophy and Religious Studies;
5001 Applied Ethics;
5002 History and Philosophy of Specific Fields;
5003 Philosophy;
5004 Religious Studies;
5005 Theology;
51 Physical Sciences;
5101 Astronomical Sciences;

- 5102 Atomic, Molecular and Optical Physics;
- 5103 Classical Physics;
- 5104 Condensed Matter Physics;
- 5105 Medical and Biological Physics;
- 5106 Nuclear and Plasma Physics;
- 5107 Particle and High Energy Physics;
- 5108 Quantum Physics;
- 5109 Space Sciences;
- 5110 Synchrotrons and Accelerators;
- 52 Psychology;
- 5201 Applied and Developmental Psychology;
- 5202 Biological Psychology;
- 5203 Clinical and Health Psychology;
- 5204 Cognitive and Computational Psychology;
- 5205 Social and Personality Psychology.

Код першої спеціальності	Код другої спеціальності	Індекс Жаккара
3504	3508	0.4849
4703	4704	0.3781
3506	3508	0.3385
3504	3506	0.3317
4008	4009	0.3216
4002	4017	0.3128
3002	3004	0.3093
3901	3903	0.2972
5106	5107	0.2937
4402	4805	0.2933
3503	3506	0.2474
4901	4904	0.2449
4007	4010	0.2363
5106	5110	0.2244
3705	3706	0.2184
3703	3706	0.2124
3703	3705	0.2108
4902	5107	0.2099
3605	4701	0.2075
4004	4011	0.2003
3802	3803	0.1992
3302	4005	0.1723
3502	3801	0.1693

4901	4903	0.1682
3501	3502	0.1669
3801	3802	0.1637
5102	5108	0.1629
5004	5005	0.1618
4902	4904	0.1589
3402	3405	0.1541
4006	4613	0.1523
4203	4205	0.1461
4604	4606	0.1441
3103	3104	0.1437
5201	5205	0.1403
4803	4807	0.1401
3304	3509	0.1389
3102	3105	0.1357
3209	5202	0.1294
3705	3709	0.1269
4407	4408	0.1266
5201	5203	0.1260
4605	4606	0.1244
3801	3803	0.1227
4006	4009	0.1225
4602	4605	0.1224
5101	5109	0.1218
3902	3903	0.1190
3204	3211	0.1177
4804	4807	0.1176
3404	3405	0.1117
3403	4016	0.1107
3406	4016	0.1100
5202	5204	0.1092
4903	4904	0.1081
4803	4804	0.1079
3702	3709	0.1074
3503	4609	0.1064
4301	4303	0.1049
5002	5003	0.1036
3107	3207	0.1008
3304	4406	0.0988
4005	4019	0.0976
4203	4206	0.0976
4804	4805	0.0970
4006	4008	0.0963
4202	4206	0.0959
4201	4207	0.0951
3507	5001	0.0940
3702	3708	0.0937
3004	3108	0.0927
3003	3109	0.0881
4602	4611	0.0878

3701	3702	0.0874
4603	4611	0.0867
4303	4705	0.0867
4006	4603	0.0855
4301	4401	0.0855
5203	5205	0.0839
3215	4204	0.0825
3603	3604	0.0810
3701	3707	0.0796
3403	3406	0.0768
4902	5106	0.0765
3701	3708	0.0763
3502	3507	0.0761
4801	4807	0.0759
3202	3211	0.0744
3903	3904	0.0736
3301	3302	0.0732
3509	4005	0.0723
3704	4013	0.0711
4901	4905	0.0703
3901	4704	0.0703
3214	5202	0.0681
3103	3109	0.0680
3707	3709	0.0653
3501	3507	0.0653
4104	4105	0.0650
3101	3404	0.0644
3204	3207	0.0643
3008	3108	0.0633
5104	5108	0.0616
4402	4804	0.0613
3701	3709	0.0610
4603	4605	0.0610
3104	3105	0.0610
4102	4104	0.0607
3301	3303	0.0601
5201	5204	0.0592
4016	4018	0.0590
4801	4806	0.0588
3502	3802	0.0582
3406	3407	0.0574
4605	4611	0.0573
4009	5102	0.0573
3004	3008	0.0573
4405	4410	0.0571
3003	3009	0.0560
3103	4102	0.0555
3107	3404	0.0543
3709	4013	0.0536
3210	4206	0.0534

3103	3108	0.0532
3201	3202	0.0532
3505	3507	0.0519
3901	3904	0.0518
4406	4410	0.0517
4704	5204	0.0514
5101	5107	0.0505
3103	4104	0.0504

ДОДАТОК В. ПЕРЕЛІК ВІТЧИЗНЯНИХ ОСВІТНІХ СПЕЦІАЛЬНОСТЕЙ

- 011 – освітні, педагогічні науки;
- 012 – дошкільна освіта;
- 013 – початкова освіта;
- 014 – середня освіта (за предметними спеціальностями);
- 015 – професійна освіта (за спеціалізаціями);
- 016 – спеціальна освіта;
- 017 – фізична культура і спорт;
- 021 – аудіовізуальне мистецтво та виробництво;
- 022 – дизайн;
- 023 – образотворче мистецтво, декоративне мистецтво, реставрація;
- 024 – хореографія;
- 025 – музичне мистецтво;
- 026 – сценічне мистецтво;
- 027 – музеєзнавство, пам'яткознавство;
- 028 – менеджмент соціокультурної діяльності;
- 029 – інформаційна, бібліотечна та архівна справа;
- 031 – релігієзнавство;
- 032 – історія та археологія;
- 033 – філософія;
- 034 – культурологія;
- 035 – філологія;
- 041 – богослов'я;
- 051 – економіка;
- 052 – політологія;
- 053 – психологія;
- 054 – соціологія;
- 061 – журналістика;
- 071 – облік і оподаткування;

- 072 – фінанси, банківська справа, страхування та фондовий ринок;
- 073 – менеджмент;
- 075 – маркетинг;
- 076 – підприємництво та торгівля;
- 081 – право;
- 091 – біологія та біохімія;
- 101 – екологія;
- 102 – хімія;
- 103 – науки про Землю;
- 104 – фізика та астрономія;
- 105 – прикладна фізика та наноматеріали;
- 106 – географія;
- 111 – математика;
- 112 – статистика;
- 113 – прикладна математика;
- 121 – інженерія програмного забезпечення;
- 122 – комп'ютерні науки;
- 123 – комп'ютерна інженерія;
- 124 – системний аналіз;
- 125 – кібербезпека та захист інформації;
- 126 – інформаційні системи та технології;
- 131 – прикладна механіка;
- 132 – матеріалознавство;
- 133 – галузеве машинобудування;
- 134 – авіаційна та ракетно-космічна техніка;
- 135 – суднобудування;
- 136 – металургія;
- 141 – електроенергетика, електротехніка та електромеханіка;
- 142 – енергетичне машинобудування;
- 143 – атомна енергетика;

- 144 – теплоенергетика;
- 145 – відновлювані джерела енергії та гідроенергетика;
- 161 – хімічні технології та інженерія;
- 162 – біотехнології та біоінженерія;
- 163 – біомедична інженерія;
- 171 – електроніка;
- 172 – електронні комунікації та радіотехніка;
- 173 – авіоніка;
- 174 – автоматизація, комп'ютерно-інтегровані технології та робототехніка;
- 175 – інформаційно-вимірювальні технології;
- 176 – мікро- та наносистемна техніка;
- 181 – харчові технології;
- 182 – технології легкої промисловості;
- 183 – технології захисту навколишнього середовища;
- 184 – гірництво;
- 185 – нафтогазова інженерія та технології;
- 186 – видавництво та поліграфія;
- 187 – деревообробні та меблеві технології;
- 191 – архітектура та містобудування;
- 192 – будівництво та цивільна інженерія;
- 193 – геодезія та землеустрій;
- 194 – гідротехнічне будівництво, водна інженерія та водні технології;
- 201 – агрономія;
- 202 – захист і карантин рослин;
- 203 – садівництво, плодоовочівництво та виноградарство;
- 204 – технологія виробництва і переробки продукції тваринництва;
- 205 – лісове господарство;
- 206 – садово-паркове господарство;
- 207 – водні біоресурси та аквакультура;

- 208 – агроінженерія;
- 211 – ветеринарна медицина;
- 221 – стоматологія;
- 222 – медицина;
- 223 – медсестринство;
- 224 – технології медичної діагностики та лікування;
- 225 – медична психологія;
- 226 – фармація, промислова фармація;
- 227 – терапія та реабілітація;
- 228 – педіатрія;
- 229 – громадське здоров'я;
- 231 – соціальна робота;
- 232 – соціальне забезпечення;
- 241 – готельно-ресторанна справа;
- 242 – туризм і рекреація;
- 251 – державна безпека;
- 252 – безпека державного кордону;
- 253 – військове управління (за видами збройних сил);
- 254 – забезпечення військ (сил);
- 255 – озброєння та військова техніка;
- 256 – національна безпека (за окремими сферами забезпечення і видами діяльності);
- 257 – управління інформаційною безпекою;
- 261 – пожежна безпека;
- 262 – правоохоронна діяльність;
- 263 – цивільна безпека;
- 271 – морський та внутрішній водний транспорт;
- 272 – авіаційний транспорт;
- 273 – залізничний транспорт;
- 274 – автомобільний транспорт;

- 275 – транспортні технології (за видами);
- 281 – публічне управління та адміністрування;
- 291 – міжнародні відносини, суспільні комунікації та регіональні студії;
- 292 – міжнародні економічні відносини;
- 293 – міжнародне право.

**ДОДАТОК Г. РЕЗУЛЬТАТИ ІДЕНТИФІКАЦІЇ СПОРІДНЕНOSTІ
ВІТЧИЗНЯНИХ ОСВІТНІХ СПЕЦІАЛЬНОСТЕЙ**

Код першої спеціальності	Код другої спеціальності	Індекс Жаккара
51	73	0.2213
122	126	0.1907
22	23	0.1667
101	183	0.1643
131	133	0.1591
104	105	0.1522
135	271	0.15
131	132	0.1477
253	255	0.1455
121	122	0.1392
31	41	0.1389
254	255	0.1311
253	254	0.1228
51	76	0.1224
81	262	0.1179
31	33	0.1176
145	194	0.1111
11	14	0.11
133	208	0.1089
132	136	0.1042
91	101	0.1023
205	206	0.0976
251	253	0.0976
171	176	0.0968
113	124	0.0952
11	13	0.095
105	171	0.0909
14	35	0.0907
103	106	0.0897
172	255	0.0896
241	242	0.0866
51	72	0.0837
11	15	0.0805
121	126	0.0803
81	293	0.0779
134	135	0.0769
102	226	0.0753
12	13	0.0745
122	123	0.0743
274	275	0.0722
51	292	0.0721
51	75	0.0714
134	173	0.0714
222	229	0.0685

17	227	0.068
143	144	0.0667
184	185	0.0667
113	122	0.0659
123	174	0.0656
52	291	0.0645
111	113	0.0645
54	231	0.0632
133	274	0.0625
222	228	0.0621
222	223	0.0612
51	71	0.0606
14	17	0.0594
21	23	0.0588
105	176	0.0588
134	272	0.0588
252	255	0.0588
11	35	0.0585
251	256	0.0556
121	124	0.0549
181	241	0.0548
73	75	0.0542
14	15	0.0535
33	52	0.0531
14	32	0.0529
23	28	0.0526
14	91	0.0514
71	72	0.0506
201	203	0.05

ДОДАТОК Д. АКТИ ВПРОВАДЖЕННЯ РЕЗУЛЬТАТІВ ДОСЛІДЖЕННЯ

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ВАСИЛЯ СТУСА



ЗАТВЕРДЖУЮ

Проректор з наукової
та навчальної роботи

Сергій РАДІО
04 2024 р.

АКТ № 08-24/12.0-24

**впровадження результатів дисертаційного дослідження
ПЕТРИЧКА Миколи Володимировича «Інформаційна
технологія підбору рецензентів для експертизи наукових
творів» у науково-організаційну діяльність Донецького
національного університету імені Василя Стуса**

1. Комісія у складі:

Голова: Олена АРТЕМЕНКОВА, в.о. завідувача науковою частиною
НДЧ, доктор філософії з соціальних комунікацій

Члени комісії: Тамара ОМЕЛЬЧУК, провідний інженер відділу розвитку
молодіжного наукового потенціалу;

Олена ГОМОН, начальник редакційного відділу навчальної і
методичної літератури;

Сергій КОЗЛОВСЬКИЙ, головний редактор журналу «Економіка
і організація управління», д-р екон. наук, професор, професор
кафедри підприємництва, корпоративної та просторової
економіки;

Роман БАБАКОВ, заступник головного редактору журналу
«Ukrainian Journal of Information Systems and Data Science», д-р
техн. наук, доцент, професор кафедри інформаційних технологій.

у період з 09.04.2024 р. по 12.04.2024 р. розглянула матеріали впровадження
результатів зазначеної дисертаційної роботи у науково-організаційну діяльність
університету.

2. Комісія встановила:

Результати дисертаційного дослідження Петричка М.В. «Інформаційна
технологія підбору рецензентів для експертизи наукових творів» за спеціальністю
126 Інформаційні системи та технології дійсно впроваджено у науково-
організаційну діяльність науково-дослідної частини Донецького національного
університету імені Василя Стуса. Впровадження здійснено у формі передачі для
дослідної експлуатації розробленої Петричком М.В. інформаційної технології

підбору рецензентів для рукописів, які надсилаються у журнали, які видає Донецький національний університет імені Василя Стуса. Пілотне використання зазначеної інформаційної технології здійснено редакціями журналів «Економіка і організація управління» та «Ukrainian Journal of Information Systems and Data Science» протягом трьох місяців. В результаті дослідної експлуатації, встановлено, що час на підбір рецензентів значно скорочується, при цьому забезпечується достатньо висока відповідність запропонованих рецензентам тематиці рукопису.

Голова комісії

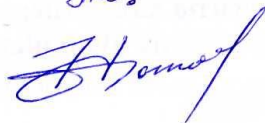


Олена АРТЕМЕНКОВА

Члени комісії:



Тамара ОМЕЛЬЧУК



Олена ГОМОН



Сергій КОЗЛОВСЬКИЙ



Роман БАБАКОВ

ЗАТВЕРДЖУЮ

Проректор з науково-педагогічної
роботи та організації освітнього
процесу Вінницького національ-
ного технічного університету



12 квітня 2024 р.

[Signature] Олександр ПЕТРОВ

АКТ

впровадження у навчальний процес підготовки магістрів
результатів PhD-дисертації Петричка Миколи
Володимировича зі спеціальності 126 «Інформаційні системи
та технології» на тему «Інформаційна технологія підбору
рецензентів для експертизи наукових творів»

Ми, що нижче підписалися, декан факультету інтелектуальних інформаційних технологій та автоматизації к.т.н., доцент В.М.Севастьянов, завідувач кафедри автоматизації та інтелектуальних інформаційних технологій, гарант освітньої-наукової програми підготовки докторів філософії за спеціальністю 126 д.т.н., професор О.В.Бісикало, професор кафедри автоматизації та інтелектуальних інформаційних технологій д.т.н., професор Р.Н. Кветний, підтверджуємо, що в навчальний процес факультету інтелектуальних інформаційних технологій та автоматизації впроваджено результати дисертаційної роботи М.В.Петричка на тему «Інформаційна технологія підбору рецензентів для експертизи наукових творів». Впровадження здійснено в рамках лекційного курсу «Методологія та організація наукових досліджень», який викладався в 2023-2024 навчальному році магістрантам спеціальності 174 «Автоматизація, комп'ютерно-інтегровані технології та робототехніка». Лектор курсу – професор Р.Н. Кветний. Основними впровадженими результатами є алгоритми та програмне забезпечення для ідентифікації рівня спорідненості наукових спеціальностей, що ґрунтується на інформаційних ресурсах системи Dimensions.

12 квітня 2024 р.

Володимир СЕВАСТЬЯНОВ

Олег БІСІКАЛО

Роман КВЕТНИЙ



МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ВІННИЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ

вул. Хмельницьке шосе, 95, м. Вінниця, 21021, тел.: (0432) 56-08-48, факс (0432) 46-57-72
 E-mail: vntu@vntu.edu.ua Код за ЄДРПОУ 02070693

05.06.24 № 15-135 на № _____ від _____

Довідка

Про участь у виконанні держбюджетної теми

Петричка Миколи Володимировича

Видана Петричку Миколі Володимировичу з підтвердженням про те, що він дійсно брав участь в якості виконавця науково-дослідної роботи 46-Д-388 «Ідентифікація прихованих залежностей в онлайнових соціальних мережах на основі методів нечіткої логіки та комп'ютерної лінгвістики», номер держреєстрації 0117U000575 у 2019р., що фінансувалась за рахунок коштів державного бюджету Міністерства освіти і науки України. Науково-дослідна робота стосувалась його дисертаційного дослідження на тему «Інформаційна технологія підбору рецензентів для експертизи наукових творів».



Проректор з наукової роботи _____

Ірина СПІФАНОВА

Начальник науково-дослідної частини _____

Олег АНТОНЮК



МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ВІННИЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ

вул. Хмельницьке шосе, 95, м. Вінниця, 21021, тел.: (0432) 56-08-48, факс (0432) 46-57-72
 E-mail: vntu@vntu.edu.ua Код за ЄДРПОУ 02070693

05.06.24 № 15-134

на №

від

Довідка

Про участь у виконанні науково-дослідної роботи

Петричка Миколи Володимировича

Видана Петричку Миколі Володимировичу з підтвердженням про те, що він дійсно брав участь в якості виконавця науково-дослідної роботи 46-K7 «Розробка методів і технологій автоматизації процесів контролю і керування» на кафедрі комп'ютерних систем управління Вінницького національного технічного університету з 2022 по 2024рр.. Науково-дослідна робота стосувалась його дисертаційного дослідження на тему «Інформаційна технологія підбору рецензентів для експертизи наукових творів».

Проректор з наукової роботи

Ірина ЄПФАНОВА

Начальник науково-дослідної частини

Олег АНТОНЮК

