

**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ імені ІГОРЯ
СІКОРСЬКОГО»**

Факультет інформатики та обчислювальної техніки
Кафедра інформаційних систем та технологій

"На правах рукопису"
УДК 004.91

ДО ЗАХИСТУ ДОПУЩЕНО
Завідувач кафедри

_____ Олександр РОЛІК

“ _____ ” _____ 2023 р.

МАГІСТЕРСЬКА ДИСЕРТАЦІЯ

на здобуття ступеня магістра

за освітньо-науковою програмою

«Інформаційні управляючі системи та технології»

зі спеціальності 126 «Інформаційні системи та технології»

на тему:

**«Підсистема валідації даних наукової діяльності наукових співробітників
закладів вищої освіти»**

Виконав:

студент VI курсу, групи ІС-01мнв

Березінський Геннадій Валерійович _____

Керівник:

Доц., к.т.н., доц. Сперкач Майя Олегівна _____

Рецензент:

Доц., к.т.н., доц. Фіногенов Олексій Дмитрович _____

Засвідчую, що у цій магістерській
дисертації немає запозичень з праць
інших авторів без відповідних
посилань.

Студент _____

Київ – 2023 року

**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»**

Факультет інформатики та обчислювальної техніки

Кафедра інформаційних систем та технологій

Рівень вищої освіти – другий (магістерський)

Спеціальність – 126 «Інформаційні системи та технології»

Освітньо-наукова програма «Інтегровані інформаційні системи»

ЗАТВЕРДЖУЮ

Завідувач кафедри

_____ Олександр РОЛІК

«__» _____ 2023 р.

**ЗАВДАННЯ
на магістерську дисертацію студенту
Березінському Геннадію Валерійовичу**

1. Тема дисертації «Підсистема валідації даних наукової діяльності наукових співробітників закладів вищої освіти»

науковий керівник дисертації доцент, к.т.н., Сперкач Майя Олегівна, затверджені
наказом по університету від « » 2023 р. №

2. Строк подання студентом дисертації “ 12 ” травня 20 23 р.

3. Об’єкт дослідження: Процес аналізу наукової діяльності наукових співробітників закладів вищої освіти.

4. Предмет дослідження: Валідація даних наукової діяльності наукових співробітників закладів вищої освіти.

5. Перелік завдань, які потрібно розробити:

Автоматизована підсистема валідації зібраних даних та виведення звіту з результатами валідації.

Графічний інтерфейс для взаємодії з системою.

Модифікація методу визначення помилкових публікацій за допомогою використання алгоритмів заснованих на графах співавторства.

Модифікація методу визначення помилкових публікацій за допомогою введення та використання коефіцієнту надійності.

6. Перелік графічного (ілюстративного) матеріалу

1. Діаграма прецедентів.
2. Повна діаграма прецедентів інформаційної системи.
3. ER-діаграма таблиць, полів та зв'язків бази даних.
4. Блок-схема попередньої версії алгоритму підтвердження авторства публікації.
5. Діаграма класів-сервісів та класів-контролерів.
6. Діаграма послідовності API-запиту на отримання кафедр факультету.
7. Діаграма послідовності запиту на отримання сторінки.

7. Перелік публікацій на тему дисертації

- Березінський Г. Аналіз публікацій викладачів закладу вищої освіти на основі відкритих даних / Г. Березінський. // Молодий вчений. – 2021. – №8. – С. 1–6.
- Березінський Г. Аналіз наукометричних даних науковців на основі графу співавторства / В. Мамонтов, Г. Березінський. // Молодий Вчений. – 2021. – №11. – С. 89–93.

8. Дата видачі завдання

“ 31 ” січня 20 23 р.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Строк виконання етапів магістерської дисертації	Примітка
1	Систематизація результатів огляду літератури	10.02	
2	Порівняльний аналіз існуючих методів аналізу авторства публікацій	20.02	
3	Визначення та опис математичних моделей	28.02	
4	Модифікація методу визначення коефіцієнту надійності	03.03	
5	Розробка інформаційного та програмного забезпечення	15.03	
7	Проведення експериментальних досліджень розроблених алгоритмів	08.04	
8	Оформлення документації	16.04	
9	Подання роботи на попередній захист	20.04	
10	Подання роботи на основний захист	17.05	

Студент

Геннадій БЕРЕЗІНСЬКИЙ

Науковий керівник

Майя СПЕРКАЧ

РЕФЕРАТ

Магістерська дисертація: 107 с., 50 рис., 15 табл., 31 джерела.

Актуальність. Актуальність даного дослідження полягає у необхідності валідації інформації про публікації наукового співробітника. Валідація полягає у вирішенні проблеми системи Google Scholar, яка виникає при встановленні дозволу на автоматичне додавання публікацій у профіль науковця.

Наявність помилково доданих публікацій в профілях наукових співробітників впливає на міжнародні рейтинги ЗВО та на показники наукових співробітників, ускладнюючи аналіз цих даних, через їх неточність.

Мета. Покращення точності визначення помилково доданих публікацій на основі метаданих. Даний метод використовує загальну інформацію про публікацію, таку як назва, автори, джерело публікації та кількість цитувань.

Для досягнення мети необхідно виконати наступні **завдання**:

- створення автоматизованої підсистеми валідації зібраних даних;
- створення графічного інтерфейсу для взаємодії з підсистемою;
- модифікація методу визначення помилкових публікацій за допомогою використання алгоритмів заснованих на графах співавторства;
- модифікація методу визначення помилкових публікацій за допомогою введення та використання коефіцієнту надійності – показника в межах від 0 до 1, яке показує ступінь коректності визначення авторства.

Об'єкт дослідження. Процес аналізу наукової діяльності наукових співробітників закладів вищої освіти.

Предмет дослідження. Валідація даних наукової діяльності наукових співробітників закладів вищої освіти.

Методи дослідження. Метод оцінки коректності визначення авторства публікації на основі метаданих публікації.

Наукова новизна. Удосконалення методу оцінки коректності визначення авторства публікації за рахунок використання коефіцієнту надійності.

Публікації: Матеріали досліджень опубліковані у:

1. Березінський Г. Аналіз публікацій викладачів закладу вищої освіти на основі відкритих даних / Г. Березінський. // Молодий вчений. – 2021. – №8. – С. 1–6.
2. Березінський Г. Аналіз наукометричних даних науковців на основі графу співавторства / В. Мамонтов, Г. Березінський. // Молодий Вчений. – 2021. – №11. – С. 89–93.

НАУКОМЕТРИЯ, GOOGLE SCHOLAR, ГРАФ, ГРАФІЧНЕ ПРЕДСТАВЛЕННЯ, ВАЛІДАЦІЯ, ЗБІР ДАНИХ, ІНДЕКС ХІРША, ПІДТВЕРДЖЕННЯ АВТОРСТВА, КРИТЕРІЙ НАДІЙНОСТІ, АНАЛІЗ АВТОРСТВА.

ABSTRACT

Master's thesis: 107 p., 50 figures, 15 tables, 31 sources.

Relevance. The relevance of this study lies in the need to validate information about a researcher's publications. Validation consists in solving the problem of the Google Scholar system, which arises when setting permission to automatically add publications to a researcher's profile.

The presence of erroneously added publications in the profiles of researchers affects the international rankings of higher education institutions and the performance of researchers, complicating the analysis of this data due to their inaccuracy.

Objective. To improve the accuracy of assessing the correctness of publication authorship based on metadata. This method uses general information about the publication, such as the title, authors, publication source, and number of citations.

To achieve the goal, the following **tasks** need to be completed:

- create an automated subsystem for validating the collected data;
- creation of a graphical interface for interaction with the subsystem;
- modification of the method for identifying false publications by using algorithms based on co-authorship graphs;
- modification of the method of identifying false publications by introducing and using the reliability coefficient, an indicator ranging from 0 to 1, which shows the degree of correctness of authorship determination.

Object of research. The process of analyzing the scientific activity of researchers of higher education institutions.

The subject of the study. Validation of data on the scientific activity of researchers of higher education institutions.

Research methods. The method of assessing the correctness of determining the authorship of a publication based on publication metadata.

Scientific novelty. Improving the method for assessing the correctness of determining the authorship of a publication by using a reliability coefficient.

Publications: Research materials are published in:

1. Berezinsky G. Analysis of publications of teachers of higher education institutions based on open data / G. Berezinsky // Young scientist. - 2021. - №8. - P. 1-6.
2. Berezinsky H. Analysis of scientometric data of scientists based on the co-authorship graph / V. Mamontov, H. Berezinsky // Young Scientist. - 2021. - №11. - P. 89-93.

SCIENTOMETRICS, GOOGLE SCHOLAR, GRAPH, GRAPHICAL REPRESENTATION, VALIDATION, DATA COLLECTION, HIRSCH INDEX, AUTHORSHIP CONFIRMATION, RELIABILITY CRITERION, AUTHORSHIP ANALYSIS.

ЗМІСТ

<u>ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ І ТЕРМІНІВ</u>	<u>9</u>
<u>ВСТУП</u>	<u>12</u>
<u>1 ПРОЄКТНІ РІШЕННЯ З РОЗРОБКИ СИСТЕМИ ЗБОРУ ТА ВАЛІДАЦІЇ ДАНИХ ПРО НАУКОВО-ПУБЛІЦИСТИЧНУ ДІЯЛЬНІСТЬ ВИКЛАДАЧІВ ЗАКЛАДУ ВИЩОЇ ОСВІТИ.....</u>	<u>15</u>
1 Опис Бізнес Процесів	15
1.1.1 Валідація даних про публікації закладу вищої освіти	15
1.1.2 Актори і функції.....	18
1.1.3 Структура бізнес-процесів	20
1.2 Опис постановки задачі	20
1.3 Рішення з інформаційного забезпечення.....	21
Висновок до розділу	26
<u>2 МОДЕЛІ ТА МЕТОДИ АНАЛІЗУ АВТОРТСВА ПУБЛІКАЦІЙ</u>	<u>27</u>
2.1 Змістовна постановка задачі.....	27
2.2 Математичні моделі.....	29
2.2.1 Розрахунок h-індексу групи науковців.....	29
2.2.2 Коефіцієнт надійності публікації	32
2.2.3 Подібність рядків	34
2.2.4 Матриці співавторства.....	35
2.3 Огляд літератури і існуючих рішень	37
2.4 Опис алгоритмів розв’язання	43
2.4.1 Перевірки на співпадіння прізвища та ініціалів.....	44
2.4.2 Перевірка джерела публікації	54
2.4.3 Детальний опис пошуку індексу Хірша для підрозділу	58
2.4.4 Побудова графу співавторства у матричному вигляді	62
2.5 Удосконалення методу оцінки надійності	65
2.6 Результати досліджень	74
Висновок до розділу	78
<u>3 РОЗРОБКА ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ.....</u>	<u>80</u>
3.1 Засоби розробки.....	80
3.1.1 SYMFONY	80

3.1.2 DOCTRINE	82
3.1.3 VUEJS	82
3.1.4 API.....	83
3.1.5 MySQL.....	84
3.1.6 BOOTSTRAP	84
3.1.7 COMPOSER	85
3.1.8 WEBPACK.....	86
3.2 АРХІТЕКТУРА ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ	87
3.3 ІНСТРУКЦІЯ КОРИСТУВАЧА.....	91
3.4 ТЕХНІЧНЕ ЗАБЕЗПЕЧЕННЯ.....	98
Висновок до розділу	100
<u>ВИСНОВКИ</u>	<u>101</u>
<u>ПЕРЕЛІК ПОСИЛАНЬ.....</u>	<u>102</u>

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

Термін	Визначення
API	Application Programing Interface. Це набір методів та інструкцій, що використовується для взаємодії різних програмних модулів однієї системи
Backend	Серверна частина застосунку. Містить в собі основну логіку роботи застосунку, встановлює зв'язок з базою даних, отримує та обробляє дані.
False-positive результат	Хибно визначене порушення валідації публікації на підтвердження авторства.
Frontend	Візуальна частина застосунку. Користувацький інтерфейс для роботи з програмою.
N-грам	Сукупність груп з n сусідніх символів, не розділених пробілом [1].
ORCID	Open Researcher and Contributor ID. Є унікальним ідентифікатором наукової роботи
REST	Representational State Transfer. Архітектурний підхід до реалізації API

UUID	Universally unique identifier. Унікальний універсальний ідентифікатор, що складається з 32 випадково згенерований 16-значних символів
Відстань Левенштейна	Відстань редагування, визначає різницю між двома рядками шляхом обчислення мінімальної кількості редагувань, які необхідні для того, щоб перетворити один рядок на інший.
ЗВО	Заклад вищої освіти
Індекс Хірша (h-індекс)	Показник, який використовується для оцінки наукового впливу окремого дослідника, визначається як кількість публікацій з порядковим номером h, така, що кількість цитувань не менша за h.
Імпакт-фактор	це числова міра, яка використовується для оцінки того, наскільки високо оцінюється науковий журнал відповідно до кількості цитувань, отриманих за певний період часу. Це значення відображає впливовість статей, опублікованих в журналі, на наукову спільноту.

Коефіцієнт надійності публікації	Значення в межах від 0 до 1, яке показує ступінь коректності авторства.
Метадані профілю	Загальна інформація профілю наукового співробітника. Ім'я, кількість цитувань, кількість публікацій, h-індекс.
Метадані публікації	Загальна інформація про публікацію. Назва, рік, автори, джерело, кількість цитувань.
Наукометрія	Дисципліна, що застосовує кількісні підходи до вивчення науки в різних її проекціях: як складного явища, як динамічного процесу та як особливої форми людської активності [2]
Регулярний вираз	Шаблон, який описує текстову послідовність.
Список Сводеша	Лінгвістичний інструмент, список зі 100 найпоширеніших слів певної мови.
Транслітерація	Відтворення кожної літери латиницею [3]

ВСТУП

З минулого століття наукова діяльність стала ключовим інструментом для пізнання світу. Протягом цього часу вона зазнала значних змін. Було прискорено процес отримання нових знань та їх впровадження у практику. Разом з цим змінилася і інфраструктура закладів науки, що повинна відповідати новим тенденціям. Це призвело до збільшення затребуваності та кількості дослідників, що займаються вузькою тематикою, напрямком чи спеціальністю. Для прискорення процесу відтворення знань та розв'язання пріоритетних проблем створюються наукові конференції, журнали та консорціуми. Тому дуже важливо, щоб наукове співтовариство мало адекватне розуміння впливу на розвиток інноваційного процесу різних груп наукових діячів, а також розуміло роль та внесок кожного дослідника у розвиток своєї галузі науки. Вирішенню даних задач сприяє оцінка наукової діяльності, заснована на наукометричних даних.

Актуальність використання наукометричних даних в галузі управління ЗВО зумовлена оцінками результативності досліджень, викладацької роботи та науково-технічного потенціалу закладу. За допомогою наукометричних показників можна визначити кількість наукових публікацій, їх цитованість, імпакт-фактор наукових видань, рейтинги науковців тощо. Ця інформація може бути використана для прийняття рішень щодо фінансування досліджень, наукового обладнання та програм навчання, оцінки ефективності наукової діяльності науковців та наукових колективів, встановлення пріоритетів у наукових дослідженнях, для рекламування наукових досягнень ЗВО, залучення фінансової та інтелектуальної підтримки від державних та приватних інституцій, а також для моніторингу науково-дослідних робіт у межах конкретних дисциплін та наукових напрямів. Додатково, наукометричні дані можуть бути корисні

для порівняння результатів досліджень з результатами інших ЗВО в Україні та за кордоном.

Системи баз даних, такі як Google Scholar[4] та Scopus[5] створені для того, щоб забезпечити зручний та швидкий доступ наукометричних показників, наукових досліджень та публікацій. Вони допомагають науковцям знайти та оцінити публікації у своїй галузі, а також забезпечують платформу для вимірювання впливу публікацій та авторів наукових праць за допомогою різних метрик. Такі системи збирають метадані про публікації, які включають автора, назву, джерело публікації, рік публікації, кількість цитувань та інші характеристики, використовуючи їх для створення профілів авторів та інституцій, а також для визначення рейтингів наукових журналів та конференцій.

Українська система вищої освіти використовує наукометричні бази даних для оцінки наукової діяльності наукових співробітників із ціллю прийняття рішень про фінансування закладів вищої освіти. Кожен заклад вищої освіти подає річний звіт про наукову діяльність своїх наукових співробітників, де вказується кількість наукових публікацій, цитування, виступів на конференціях тощо. Ці дані перевіряються науковими експертами та аналізуються за допомогою наукометричних інструментів, таких як Google Scholar та Scopus. На основі цих даних приймаються рішення про виділення фінансування тим закладам вищої освіти, які мають найбільшу наукову продуктивність та внесок у розвиток науки в Україні [6].

Актуальність розробки полягає у необхідності валідації інформації про публікації наукового співробітника. Валідація полягає у вирішенні проблеми системи Google Scholar, яка виникає при встановленні дозволу на автоматичне додавання публікацій у профіль науковця. Виконуючи пошук по кожному науковцю у відкритих джерелах, Google Scholar автоматично додає у свою базу ті публікації, автором яких вважає даного науковця. У

більшості випадків, такі публікації додаються коректно. Проте іноді до профілю наукового співробітника публікація додається помилково.

Метою даної роботи є покращення точності визначення помилково доданих публікацій на основі метаданих. Даний метод використовує загальну інформацію про публікацію, таку як назва, автори, джерело публікації та кількість цитувань. Метод покращує попередню версію [7], який мав велику кількість false-positive результатів та низьку швидкість роботи. На основі даної системи проведено дослідження ефективності різних підходів та методів визначення помилкових публікацій. Для демонстрації роботи використано список наукових співробітників Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського».

Об'єкт дослідження. Процес аналізу наукової діяльності наукових співробітників закладів вищої освіти.

Предмет дослідження. Валідація даних наукової діяльності наукових співробітників закладів вищої освіти.

1 ПРОЄКТНІ РІШЕННЯ З РОЗРОБКИ СИСТЕМИ ЗБОРУ ТА ВАЛІДАЦІЇ ДАНИХ ПРО НАУКОВО-ПУБЛІЦИСТИЧНУ ДІЯЛЬНІСТЬ ВИКЛАДАЧІВ ЗАКЛАДУ ВИЩОЇ ОСВІТИ

1 Опис Бізнес процесів

В даному розділі наводиться опис бізнес-процесів, які описують актуальність проведених досліджень, наводяться діаграми використання та послідовностей.

1.1.1 Валідація даних про публікації закладу вищої освіти

Основним недоліком використання сервісу Google Scholar може бути автоматичне додавання публікацій в профіль наукового співробітника. Хоч система Google Scholar є більш функціональною за базу ORCID, проте у ORCID, навіть за однакових прізвища й імені, не буде співпадінь авторів та додавання чужих публікацій, як це досить часто відбувається в Google Scholar (якщо в останньому обрано автоматичне оновлення статей) [8]. Це є суттєвим недоліком та може призводити до проблем при будівництві рейтингів, заснованих на даних системи Google Scholar, ускладнює аналіз цих даних через потенційну наявність хибних даних у профілі, які можуть впливати на загальні наукометричні показники наукового співробітника закладу вищої освіти та показники наукових підрозділів закладів вищої освіти в цілому.

Наявність помилково доданих публікацій в профілях наукових співробітників при підрахунку міжнародних рейтингів закладів вищої освіти (Webometrics[9], UniRank[10], QS[11] та ін.) карається штрафними санкціями: штрафні рейтингові бали, або вилучення навчального закладу з

рейтингу при великій кількості порушень. Тому своєчасне визначення помилково доданих публікацій є важливим [7].

Таким чином, є необхідність мати змогу автоматизовано оцінити коректність додання публікації в профіль на основі метаданих цієї ж публікації, так як не завжди є можливість оцінити і перевірити сам текст публікації. Втім, перевірка тексту публікації не може бути автоматизована без використання методів машинного навчання.

При ручному додаванні публікацій також можливі виникнення неточностей. Користувачі наукометричних баз даних можуть вказувати помилкові дані при заповненні метаданих публікацій – робити похибки при вказуванні авторів, вказувати не актуальні дані (у варіанті зміни прізвища) або мати не актуальні дані про себе у профілі.

Також варто враховувати можливість публікації іноземними мовами – тоді можуть помилки транслітерації при заповненні авторства. Проблемою транслітерації є різні варіанти використання одного імені при початковому перекладі з різних мов, відсутність єдиного чітко визначеного стандарту транслітерації та недотримання існуючих рекомендованих стандартів для транслітерації імен з української мови на іноземні.

Суттєвою проблемою також є питання актуальності даних. Наукові співробітники можуть змінювати наукові підрозділи всередині одного навчального закладу, змінювати навчальні заклади або призупиняти професійну діяльність. Важливо враховувати, що їх профілі у наукометричних базах даних мають відповідно актуалізуватись, для того щоб вони не впливали на показники навчального закладу або наукового підрозділу, до якого вони більше не належать.

Відсутність стандартизації у наукометричних базах є іншою проблемою, за якою ЗВО мають слідкувати. Так, у наукових співробітників може бути декілька профілів у одній наукометричній базі, профілі які не містять актуальних даних, нові профілі, що виникають у разі зміни

прізвища, що може ускладнити співставлення публікації та її автора та інше.

Питання єдиної ідентифікації є однією з не вирішених на сьогоднішній день проблем. Відсутність визначеного єдиного ідентифікатора наукового співробітника, що міг би використовуватись для полегшеного співставлення наукометричного профілю, науковця та публікації ускладнює процес аналізу науково-публіцистичної діяльності наукових співробітників ЗВО. Одним з таких ідентифікаторів міг би виступати ORCID, але він не є широко поширеним та багато профілів науковців не зв'язані з відповідним ORCID профілем. Більш широке використання ORCID дозволило би вирішити питання неоднозначності при аналізі профілів в наукометричних базах, так як вирішило би питання повних однофамільців, різних варіантів написання імені, транслітерації на інші мови, зміну імені чи прізвища, тощо.

Описані проблеми можуть призвести до помилкових результатів при оцінці підтвердження авторства публікації, ускладнити аналіз наукометричних даних або призводити до неточностей при пошуку інформації про науково-публіцистичну діяльність наукових співробітників ЗВО.

Отже необхідно визначити коректний метод для оцінки підтвердження авторства публікації, розробити алгоритм, що реалізовує метод підтвердження авторства, реалізувати інформаційну систему, яка буде використовувати розроблений алгоритм та представити розроблений метод для оцінки коректності визначення авторства публікації з системи Google Scholar. За допомогою користувацького інтерфейсу, підсистема надаватиме можливість перегляду інформації про факультети, кафедри, наукових співробітників та їх публікації, перегляди інформацію про авторів публікації та запускати перевірку конкретних публікацій.

Так як метод засновано на метаданих публікацій, ті з них, в яких наявна неоднозначність з інформацією в профілі будуть також позначатись як потенційно хибні. Це дозволить звернути увагу на публікації, в описі яких допущено помилки, публікації, в яких правило транслітерації авторів відрізняється від правила транслітерації профілю та ін.

1.1.2 Актори і функції

Дане проєктне рішення пропонує підсистему валідації даних у складі інформаційної системи для збору, валідації та аналізу даних про науково-публіцистичну діяльність наукових співробітників закладу вищої освіти. Функції даної підсистеми – перегляд даних про факультети кафедри, окремих наукових співробітників та публікації, а також валідація публікацій, взятих з профілів Google Scholar наукових співробітників ЗВО.

Дана підсистема має в собі двох акторів – користувач та системний адміністратор.

Користувач має доступ до веб-інтерфейсу системи, та може переглядати інформацію про факультети, кафедри або сторінки наукового співробітника з їх списком публікацій. Користувач повинен мати можливість виконати пошук за будь-яким з параметрів одночасно: назвою, роком, авторами та видавництвом. Користувач повинен мати можливість переглядати опис публікацій, перейшовши до неї зі списку публікацій наукового співробітника, або пошуку.

Адміністратор може запускати скрипт для валідації усіх наявних в базі публікацій. Цей процес є довгим та запускається за допомогою .sh скрипта. Даний скрипт зробить перевірки для усіх наявних публікацій в системі та оновить їх значення коефіцієнту надійності.

Описані вище актори та функції можуть бути представлені у вигляді UML діаграми на рис. 1.1.

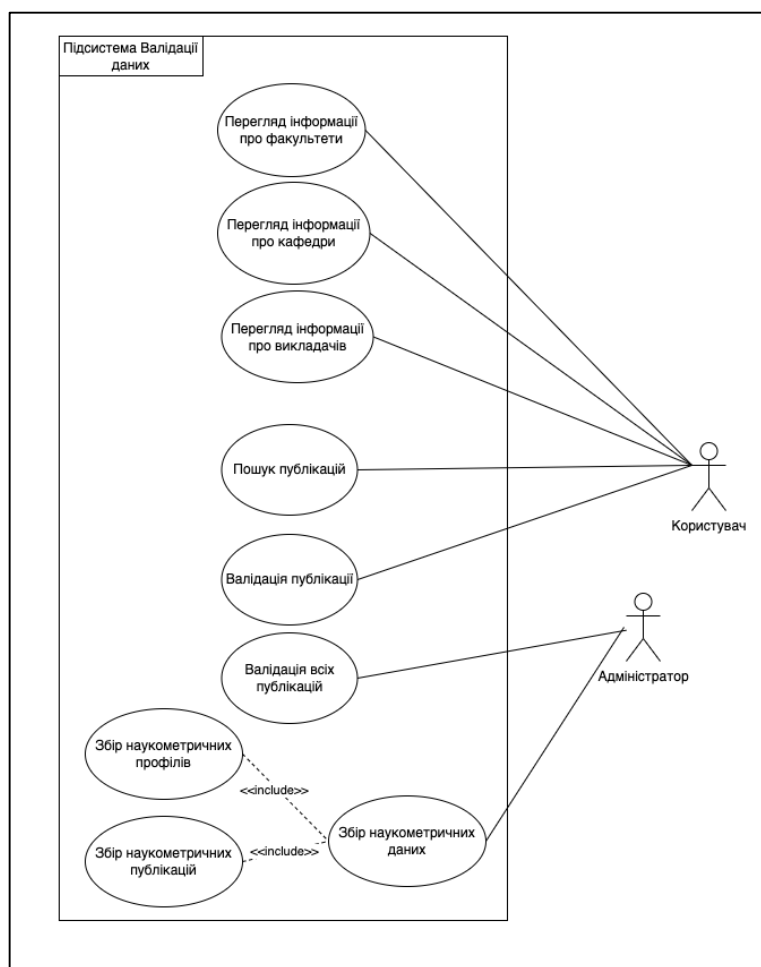


Рисунок 1.1 – Діаграма прецедентів

Даний рисунок демонструє акторів зазначеної підсистеми та функції, які вони мають у вигляді use-case діаграми прецедентів.

Зазначена підсистема є складовою у системі збору, валідації та аналізу даних. Система розширює основні функції підсистеми до проведення аналізу отриманих даних, аналізу наукометричних показників, використання графів співавторства для дослідження зв'язків між науковими співробітниками закладу вищої освіти та відображення даних для подальшого аналізу. Повна система у вигляді діаграми прецедентів представлена у «Додатку А» на «Плакати 1».

1.1.3 Структура бізнес-процесів

Основний бізнес-процес даної підсистеми – валідація даних про публікації наукових співробітників ЗВО. Процес валідації здійснюється за допомогою .sh скрипту для валідації усіх публікацій, або відповідної кнопки на сторінці публікацій наукового співробітника. Коротко даний процес представлений в «Додатку А» на «Плакаті 2».

Важливими елементами загальної системи є використання та отримання даних. Отриманням даних є автоматизований процес збору наукометричних даних з відкритих джерел Google Scholar та Scopus. Використання даних полягає у аналізі наукометричних показників та зв'язків між науковими співробітниками. Етап валідації даних є важливим через те, що він дозволяє використовувати для аналізу більш коректні дані, не брати до уваги помилково додані публікації, які можуть суттєво вплинути на показники наукового співробітника, або наукового підрозділу загалом.

1.2 Опис постановки задачі

Одним з важливих показників розвитку наукових напрямлень закладу вищої освіти є наукометричні показники, такі як загальна кількість публікацій, показник цитованості та наукометричний індекс (у рамках даного дослідження розглядається в основному h-index). Дані показники визначають вплив наукової діяльності на показники закладу вищої освіти при складанні рейтингів, розподілі фінансування та процесів щодо покращення організації публіцистичної діяльності наукових співробітників закладу вищої освіти.

Призначення розробки - створення автоматизованої системи для допомоги у зборі даних про науково-публікаційну діяльність науковими

співробітниками закладів вищої освіти на основі вхідних даних у вигляді списку наукових співробітників закладу вищої освіти. Підсистема валідації також допомагає визначати помилково додані публікації сервісом Google Scholar за допомогою оцінки коректності визначення авторства публікації.

Дана підсистема у складі системи збору, валідації та аналізу даних дозволить автоматизувати процес збору даних про науково-публікаційну діяльність наукових співробітників закладу вищої освіти та автоматизувати процес пошуку та визначення помилково доданих публікацій у сервісі Google Scholar.

Підсистема дозволяє проводити валідацію окремих публікацій або запускати процес валідації усіх публікацій у базі даних.

Підсистема має графічний інтерфейсу для взаємодії з системою. За допомогою графічного інтерфейсу необхідно відобразити зібрані дані та результати валідації. Графічний інтерфейс дозволяє переглядати інформацію про наукові підрозділи та наукових співробітників.

1.3 Рішення з інформаційного забезпечення

Зібрані дані зберігатимуться у реляційній базі даних MySQL, що дозволяє побудувати чіткі зв'язки між сутностями-таблицями системи, простіше взаємодіяти з системою, використовуючи високорівневі запити через REST API. ER-діаграма бази даних, що показує таблиці-сутності та поля бази даних представлена в «Додатку А» на «Плакати 3». В таблицях 1.1-1.7 представлено опис полів бази даних.

Таблиця 1.1 – Таблиця бази даних professor

№	Назва поля	Тип	Призначення
1	id	varchar(36)	Ідентифікатор запису в таблиці

2	name	varchar(128)	Ім'я наукового співробітника закладу вищої освіти
3	faculty	varchar(128)	Коротка назва факультету наукового співробітника закладу вищої освіти
4	cathedra	varchar(1280)	повна назва кафедри наукового співробітника закладу вищої освіти
5	scholar_link	varchar(64)	посилання на профіль в системі Google Scholar
6	scopus_link	varchar(64)	посилання на профіль у системі Scopus
7	errors	varchar(1024)	знайдені помилки профілю у системі Google Scholar
8	is_active	varchar(1)	булева змінна що позначає чи активний профіль наукового співробітника
9	scholar_profile_id	varchar(36)	унікальний uuid ідентифікатор профілю системі Google Scholar наукового співробітника закладу вищої освіти
10	scopus_profile_id	varchar(36)	Унікальний uuid ідентифікатор профілю в системі Scopus наукового співробітника закладу вищої освіти

Таблиця 1.1 зберігає формалізовану вхідну інформацію про наукового співробітника закладу вищої освіти. Містить дані про факультет, кафедру та посилання на профілі в системах Google Scholar та Scopus, якщо вони наявні.

Таблиця 1.2 – Таблиця scholar_profile

№	Назва поля	Тип	Призначення
1	id	varchar(36)	Унікальний ідентифікатор запису в таблиці
2	name_in_system	varchar(512)	Ім'я зазначена в профілі
3	total_docs	int	Загальна кількість публікацій
4	total_docs_last_5_years	int	Кількість публікацій за останні 5 років
5	total_citation	int	Загальна кількість цитувань
6	total_citation_last_5_years	int	Загальна кількість цитувань за останні 5 років
7	h_index	int	Індекс Хірша наукового співробітника
8	h_index_last_5_years	int	Індекс Хірша на основі 5 років
9	professor	varchar(36)	Унікальний ідентифікатор наукового співробітника

Таблиця 1.2 містить загальні дані про профіль наукового співробітника в профілі Google Scholar.

Таблиця 1.3 – Таблиця scholar_profile_publication

№	Назва поля	Тип	Призначення
1	id	varchar(36)	Ідентифікатор запису в таблиці
2	profile	varchar(36)	Ідентифікатор профілю наукового співробітника

3	publication	varchar(36)	Ідентифікатор публікації
4	is_valid	int	Рішення про валідність публікації
5	reason	varchar(256)	Причина рішення про валідність
6	validity	double	Коефіцієнт надійності публікації
7	comment	varchar(256)	Коментар до рішення про коефіцієнт надійності публікації

Таблиця 1.3 є представленням зв'язку багато до багатьох між профілем та публікацією в систем Google Scholar.

Таблиця 1.4 – Таблиця scholar_publication

№	Назва поля	Тип	Призначення
1	id	varchar(36)	Унікальний ідентифікатор запису
2	name	varchar(2560)	Назва публікації з системи Google Scholar
3	authors	varchar(2800)	Автор публікації
4	year	varchar(8)	Рік публікації
5	source	varchar(600)	Джерело публікації
6	citations	int	Кількість цитувань

Таблиця 1.4 містить загальну інформацію про унікальну публікацію в системі Google Scholar.

Таблиця 1.5 – Таблиця scopus_profile

№	Назва поля	Тип	Призначення
1	id	varchar(36)	Унікальний ідентифікатор запису

2	professor	varchar(36)	Посилання на ідентифікатор наукового співробітника
3	orcid	varchar(64)	orcid ідентифікатор
4	total_docs	varchar(11)	загальна кількість публікацій
5	total_citations	varchar(11)	загальна Кількість цитувань
6	h_index	varchar(11)	Індекс Хірша в системі Scopus

Таблиця 1.5 містить загальну інформацію про профіль у системі Scopus.

Таблиця 1.6 – Таблиця scopus_profile_publication

№	Назва поля	Тип	Призначення
1	id	varchar(36)	Унікальний ідентифікатор запису
2	publication	varchar(36)	Ідентифікатор публікації
3	profile	varchar(36)	Ідентифікатор профілю

Таблиця 1.6 є представленням зв'язку багато до багатьох між профілем та публікацією в системі Scopus.

Таблиця 1.7 – Таблиця scopus_publications

№	Назва поля	Тип	Призначення
1	id	varchar(36)	Унікальний ідентифікатор запису
2	name	varchar(2560)	Назва публікації
3	authors	varchar(2800)	Автори публікації
4	year	varchar(8)	Рік публікації
5	source	varchar(600)	Джерело публікації

6	citations	int	Кількість цитувань
---	-----------	-----	--------------------

Таблиця 1.7 містить загальну інформацію про унікальну Scopus публікацію

Таблиці `doctrine_migration_versions` та `messenger_messages` є внутрішніми таблицями та використовуються фреймворками Doctrine та Symfony для забезпечення своєї роботи. Дані фреймворки будуть більш детально писані у Розділі 3 даної дисертації.

Висновок до розділу

Даний розділ представляє опис проєктних рішень для магістерської дисертації. Розділ описує бізнес-процеси, наводячи їх формальний опис та UML діаграму прецедентів. Розділ визначає мету та список задач, які необхідно реалізувати згідно описаних бізнес-процесів.

Розділ описує рішення з інформаційного забезпечення наводячи ER-діаграму бази даних та формальний опис таблиць та зв'язків бази даних.

2 МОДЕЛІ ТА МЕТОДИ АНАЛІЗУ АВТОРТСВА ПУБЛІКАЦІЙ

Даний розділ представляє опис математичних моделей та алгоритмів, які використовуються для досягнення поставленої мети магістерської дисертації. У розділі оглянуто літературу, що наводить існуючі дослідження, які були розглянуті при формуванні основного алгоритму для роботи в рамках дисертації. Даний розділ представляє метод розрахунку коефіцієнту надійності публікації, наводить особливості його розробки та аналізу, а також результати роботи методу, які задовольняють поставлену мету даної дисертації.

2.1 Змістовна постановка задачі

Задачу дисертації даних можна представити у вигляді двох підзадач – збору даних та їх валідації.

Першою з задач інформаційної системи є автоматизоване отримання даних з відкритих джерел. Для цього використано метод автоматизованого синтаксичного аналізу веб ресурсу (парсинг).

В загальному випадку процес парсингу можна розбити на наступні етапи:

Етап 1. Пошук даних – у програму-парсер завантажується вихідний HTML-код сторінки сайту. З кодом починає працювати скрипт, який розбиває текст на лексеми, виділяючи необхідну інформацію.

Етап 2. Вилучення інформації – відбувається пошук даних за визначеним шаблоном для подальшого збереження.

Етап 3. Збереження даних – після отримання інформація зберігається у вигляді таблиць або вноситься до бази даних.

Задача збору даних полягає у автоматизованому пошуку та збереженні даних. Для пошуку використовуються стандартизовані Excel таблиці, які містять інформацію про наукового співробітника – його прізвище, ім'я, по-батькові, факультет, кафедру та посилання на Google Scholar та Scopus профілі.

За допомогою отриманих посилань, автоматизований синтаксичний аналізатор збирає інформацію про наукового співробітника з його профілю – прізвище, ім'я, по-батькові, кількість публікацій й цитувань, індекс Хірша. Для сервісу Google Scholar додатково збираються значення кількості публікацій, цитувань, а також індекс Хірша за останні 5 років для більш точного дослідження поточного стану публіцистичної активності.

Зібрана інформація про науково публіцистичну діяльність має бути перевірена перед подальшим використанням по визначених правилах, які підтверджують коректність даних. У загальному розумінні валідація – підтвердження відповідності, описаний набір процедур, що дозволяють визначити правильність і корисність даних. Валідація використовує підпрограми, які часто називають «правилами перевірки», «обмеженнями перевірки» або «програмами перевірки», які перевіряють правильність, змістовність і безпеку даних, що вводяться в систему. Правила можуть бути реалізовані за допомогою автоматизованих засобів словника даних, або шляхом включення явної логіки перевірки прикладної програми комп'ютера та його застосування.

Це відрізняється від формальної перевірки, яка намагається довести або спростувати правильність алгоритмів для реалізації специфікації або властивості.

Валідація помилково доданих публікацій вже була описана у попередніх дослідженнях [7]. Дане дослідження ставить на меті

оптимізацію існуючого алгоритму за допомогою використання графів співавторства для більш швидкого пошуку співавторів публікації, а також зменшення алгоритмічної складності і часу роботи валідатора.

Другим варіантом покращення є введення коефіцієнту надійності публікації, та заміна 0-1 методу на метод, заснований на коефіцієнті надійності. Також, додаючи мінорні перевірки, як-от чи публікувався даний науковець в певному журналі раніше, чи є категорія публікації релевантною (вже зустрічались) у наукового співробітника, кількість співавторів публікації тощо. Дані критерії не можна було використовувати у 0-1 алгоритмі через недостатність виконання однієї з умов для того, щоб визнати публікацію потенційно помилково доданою.

Публікації додаються у систему, створюючи зв'язок багато-до-багатьох, таким чином, що кожен науковий співробітник, що є її автором, матиме зв'язок з нею, а сама публікація матиме зв'язок з усіма авторами.

2.2 Математичні моделі

Для реалізація поставлених задач, необхідно реалізувати методи перерахунку індексу Хірша для наукового підрозділу (групи наукових співробітників), задачу вирішення подібності рядків, побудови матричного графу співавторства, і основного методу дисертації – обчислення коефіцієнту надійності публікації. Дані методи можуть бути представлені за допомогою математичних моделей. Математичні моделі для цих методів представлено нижче.

2.2.1 Розрахунок h-індексу групи науковців

H-індекс - це показник, який використовується для оцінки наукового впливу окремого дослідника. Однак він також може бути застосований до

групи авторів, щоб виміряти їхній колективний вплив. Метод розрахунку h-індексу для групи авторів:

Крок 1. Необхідно зібрати всі публікації групи авторів, включаючи журнальні статті, доповіді на конференціях, розділи книг та інші наукові роботи.

Крок 2. Визначити кількість цитувань для кожної публікації; цю інформацію зазвичай можна знайти в таких базах даних, як Web of Science, Google Scholar або Scopus.

Крок 3. Відсортувати публікації в порядку зменшення кількості цитувань, які вони отримали.

Крок 4. Починаючи з верхньої частини списку, вирахувати кількість публікацій, які мають принаймні стільки ж цитувань, скільки їхній рейтинг. Наприклад, h-індекс групи авторів з 10 публікаціями дорівнює 6, якщо шоста публікація має щонайменше шість цитувань.

Крок 5. Отримане число i є h-індексом групи авторів.

Необхідно враховувати, що h-індекс – це лише один із способів оцінити вплив групи авторів. Важливо враховувати інші фактори, такі як якість і значущість досліджень, при оцінці загального впливу.

Представимо формальний опис у вигляді математичної моделі. Припустимо, що у нас є група авторів з N публікаціями, і назвемо кількість цитувань для кожної публікації C_i . Ми можемо визначити h-індекс (h) як максимальне значення i , при якому група має i публікацій з принаймні i цитуваннями.

Представимо наведений вище метод обчислення h-індексу у формі математичної моделі:

Крок 1. Сортування списку публікацій за цитування за спаданням, наведено у формулі 1 нижче.

$$C_1 \geq C_2 \geq \dots \geq C_N \quad (1)$$

Крок 2. Підрахунок кількості публікацій, які мають принаймні стільки ж цитувань, як їх порядковий номер, починаючи з вершини списку, у формулі 2.

$$h = \max \{ i \mid C_i \geq i \} \quad (2)$$

Крок 3. Якщо немає публікацій, які мають щонайменше h цитувань, тоді h вважається 0.

Ця модель передбачає, що є відомий список публікацій, і для кожної публікації вже відома кількість цитувань.

Наприклад, розглянемо групу авторів з такими показниками цитованості публікацій:

$$C_1 = 10, C_2 = 8, C_3 = 5, C_4 = 4, C_5 = 4, C_6 = 3, C_7 = 2, C_8 = 1$$

Відсортований список цитувань, згідно (1), має вигляд:

$$C_1 \geq C_2 \geq C_3 \geq C_4 \geq C_5 \geq C_6 \geq C_7 \geq C_8$$

Починаючи з верхньої частини списку, ми можемо підрахувати кількість публікацій, які мають принаймні стільки ж цитувань, як їх номер у відсортованому списку, відповідно (2):

$$h = \max \{ i \mid C_i \geq i \} = \max \{ 1, 2, 3, 4, 4, 5, 6, 0 \} = 4$$

Отже, h -індекс для цієї групи авторів дорівнює 4.

2.2.2 Коефіцієнт надійності публікації

Для методу визначення коректності авторства публікації вводиться список критеріїв, вага кожного з них, та коефіцієнт довіри до публікації наукового співробітника. Коефіцієнт довіри – значення в межах від 0 до 1, яке показує ступінь коректності визначення авторства. Низький показний коефіцієнту довіри показує, що, потенційно, публікація була додана помилково. За замовчуванням, у всіх публікацій коефіцієнт довіри рівний 1. Тобто на момент збереження зібраних даних, усі публікації є достовірними та релевантними. На етапі валідації, кожна публікація проходить через перевірки, які, у випадку провалу валідації перевіркою, зменшують показник коефіцієнту довіри публікації відповідно до ваги критерію.

Так, значимі критерії, як, наприклад, неспівпадіння прізвища, матимуть більш вагомий ефект. В той час як незначні критерії, як-от відсутність співавтора в межах досліджуваного ЗВО, що раніше використовувались лише в комплексі з іншими перевітками, не матимуть суттєвого впливу на коефіцієнт довіри.

Розглянемо критерії та їх вплив на коефіцієнт довіри. Планується провести декілька спроб валідації для визначення показника впливу (коефіцієнтів) обраних критеріїв. Коригування буде проводитись для досягнення максимально точного результату та знаходження оптимальних коефіцієнтів впливу параметрів на показник коефіцієнту довіри. Також для досягнення більшої точності результатів, деякі критерії в процесі налаштування можуть бути замінені на інші.

Визначені критерії отримали початкові значення коефіцієнтів впливу з залежності від значимості критерію, необхідності його обов'язкового виконання (як співпадіння прізвище) або визначення критерію, як допоміжного з низьким впливом, як перевірка співавторів, так як це не є

надійною перевіркою. Початкові значення коефіцієнтів впливу критеріїв представлені у таблиці 2.1.

Проходження критеріїв оцінки коефіцієнту довіри буде ітераційним – тобто тестування критеріїв відбуватиметься у описаному порядку, та буде спиратися на результати попередніх критеріїв. Початкове значення критерію – 1, це означає повну довіру до публікації, і що вона на 99% є авторства вказаного науковця (1% списується на статистичну похибку, можливі похибки обчислень та невраховані випадки).

Поточне значення коефіцієнту довіри на ітерації прийнято за x для спрощення.

Таблиця 2.1 – Початкові критерії перевірки та коефіцієнти їх впливу

№	Критерій	Коефіцієнт впливу
1	Неспівпадіння прізвища	$0.25x$
2	Неспівпадіння ініціалів	$0.6x$
3	Відсутність співавтора у КПІ	$0.8x$
4	Відсутність публікацій у даному журналі	$0.75x$
5	Кількість цитувань публікації відрізняється від середньої, без врахування цієї більш ніж на 50%	$0.9x$
6	Індекс Хірша без врахування даної публікації відрізняється від індексу Хірша більш ніж на 50%	$0.8x$
7	Кількість зв'язків у вершини графу співавторства науковця, де n – кількість пов'язаних нод. При цьому максимальне значення n не перевищує 10	$0.05 * 2nx$

Таким чином, впродовж першого етапу модифікації методу, значення коефіцієнту довіри може змінюватись від 0.0648 до 1. При цьому можна зробити припущення, що при значенні коефіцієнту <0.3 , таку публікацію можна вважати помилково доданою; у межах $[0.3, 0.8]$ – публікація потенційно не надійна; при значенні коефіцієнту довіри >0.8 – публікація надійна.

Важливим є уточнення, що дані критерії та їх вплив на коефіцієнт довіри можуть змінюватись у процесі налаштування значень критеріїв.

Таким чином, очікується за 3-4 ітерації експерименту вивести оптимальні значення впливу критеріїв та значення коефіцієнту довіри до публікації. Враховуючи збільшену кількість критеріїв, необхідність постійного перерахунку середньої кількості цитувань та індексу Хірша, час роботи алгоритму, заснованому на даному методі, може бути збільшений, але не суттєво, через спрощення роботи з співавторами завдяки графу співавторства. Водночас це дасть більшу точність за рахунок можливості використання діапазону довіри до публікації, а не бінарного значення 0-1, що може бути помилковим через зазначені вище помилки у веденні профілю та оформленні публікацій у сервісі Google Scholar.

Після завершення експериментальних прогонів кожна публікація буде оцінена за кінцевими критеріями та їх коефіцієнтами, для них буде визначений коефіцієнт довіри, що буде виводитись в інтерфейсі користувача в інформації про публікації наукового співробітника.

2.2.3 Подібність рядків

Для порівняння джерел публікації та співставлення авторів з профілем буде використовуватись алгоритм знаходження відстані між рядками – відстань Левенштейна. Відстань Левенштейна, також відома як відстань редагування, визначає різницю між двома рядками шляхом обчислення

мінімальної кількості редагувань, які необхідні для того, щоб перетворити один рядок на інший. Ці редагування можуть бути вставками, вилученнями або замінами одного символу на інший.

Математична модель для відстані Левенштейна представлена наступним чином.

Нехай є два рядки з довжиною n та m , *ЖУРНАЛ1* та *ЖУРНАЛ2*. відстань Левенштейна між цими рядками буде представлена функцією $d(i,j)$, що вираховує мінімальну кількість односимвольних редагувань (вставка, вилучення, заміна), що необхідні для перетворення підрядка $(0,i)$ рядка *ЖУРНАЛ1* в підрядок $(0,j)$ рядка *ЖУРНАЛ2*. функція $d(i,j)$ задається рекурсивною формулою, яку можна представити у вигляді псевдокоду наступним чином:

```

d(i, j) = {
i, якщо j = 0
j, якщо i = 0
d(i-1, j-1), якщо string1[i-1] = string2[j-1]
1 + min(d(i-1, j), d(i, j-1), d(i-1, j-1)),
інакше
}

```

Наведений метод обчислює відстань Левенштейна для заданих рядків.

2.2.4 Матриці співавторства

Як було зазначено в попередніх дослідженнях [12], одним з критеріїв для визначення коефіцієнту надійності публікації є рішення, що приймається на графі співавторства. Розглянемо визначення графових

структур. Граф є гнучкою та універсальною структурою даних, що використовується в інформаційних технологіях.

Дискретна математика описує графи як колекцію нод (вершин) та ребер, які з'єднують між собою вершини. Так, множина вершин не може бути порожньою, в той час як множина ребер – може [13]. Графи часто використовуються для моделювання відношення між об'єктами-вершинами, такими як елементи в комп'ютерній мережі, транспортні зв'язки між пунктами призначення або, в даному випадку, для демонстрації зв'язків між науковими співробітниками ЗВО. Вершини графу представляють модельовані елементи: комп'ютери в спільній мережі, населені пункти тощо; зв'язки між ними – елементи, що показують відношення – складові, що поєднують комп'ютери в мережу, дороги між містами, або публікації зі спільним авторством пар авторів між собою.

Графи поділяються на орієнтовані та неорієнтовані. В неорієнтованих графах ребра не мають напрямку, і представляють собою двосторонній зв'язок між об'єктами, що може бути пройдений в обох напрямках.

За ребрами графи поділяються на зважені та незважені. В зважених графах ребра мають вагу – величину, що показує силу відношення між парою вершин графів – наприклад, відстань між пунктами призначення, або кількість спільних публікацій між науковими співробітниками.

Графи мають багато застосувань в комп'ютерних науках, зокрема в алгоритмах пошуку найкоротших шляхів, мережових потоків і зіставлення. Вони також використовуються в аналізі соціальних мереж, машинному навчанні та багатьох інших сферах.

Розглянемо детальніше граф співавторства, який буде представляти зв'язки між науковими співробітниками ЗВО. Для простоти будемо зображати такий граф за допомогою матриці $n \times n$, де n - множина наукових співробітників, а елемент a_{ij} буде показувати кількість спільних публікацій

авторів i та j з множини n . Матриця розширюється до розміру $m \times m$, де m - загальна співавторів та їх співавторів.

Формально математична модель даного графу представлена матрицею:

$$\begin{matrix} a_{11} & a_{12} & \dots & a_{1n} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2n} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} & \dots & a_{nm} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} & \dots & a_{mm} \end{matrix},$$

де елементом a_{ij} буде кількість спільних публікацій у наукового співробітника i та j .

Якщо елемент a_{ij} рівний 0, то це означає, що ребра між заданими вершинами не має, і наукові співробітники не мають спільних публікацій між собою.

Матричне представлення графу співавторства використовується для більш спрощеного способу відображення графу у frontend частині застосунку для подальшої візуалізації.

2.3 Огляд літератури і існуючих рішень

Даний розділ розглядає основні публікації та дослідження на тематику даної задачі.

Дослідження [14] описує загальні засади використання індексу Хірша, його методологію та похідні. В рамках даної магістерської дисертації, дослідження [14] використовується для визначення впливу індексу Хірша на загальні показники публіцистичної активності наукових підрозділів з метою визначення оптимального показника для демонстрації наукової активності.

Дослідження [15] пропонує математичні методи оцінювання науково-публіцистичної діяльності групи науковців, розглядаючи їх як єдину неподільну одиницю. Згідно загаданої роботи, індекс Хірша підрозділу визначатиметься наступними кроками:

Крок 1. Упорядкування публікацій за кількістю цитувань n по спаданню.

Крок 2. Знаходження публікації з найбільшим порядковим номером, такої, що кількість цитувань буде прямо пропорційною порядковому номеру.

Крок 3. Привласнення значення порядкового номеру публікації індексу Хірша.

Втім, така методологія має недоліки – вона унеможлиблює коректне порівняння наукових підрозділів різної спеціалізації. Так, наприклад, біологічні науки цитуються менше за економічні. Користь даного дослідження полягає у приведенні алгоритму розрахунку індексу Хірша для окремого науковця, що може також використовуватись для розрахунку індексу Хірша для наукового підрозділу. Це корисно для побудови звітів по підрозділам та окремим науковим групам [15].

Робота [16] описує поняття метаданих наукової публікації, ключові характеристики опису наукової роботи, та наводить формальну модель метаданих публікації.

З огляду на наведені дослідження, були визначені ключові критерії, на основі яких будувались перевірки для підтвердження авторства публікації.

Дослідження [17] описує загальні правила ведення профілів Google Scholar. В даній публікації наведена як проблема автоматичного додавання, так і загальні можливі проблеми профілів, наряду з загальними рекомендаціями.

Робота [7] є попереднім дослідженням автора цієї дисертації, що представляє початкову версію алгоритму підтвердження авторства публікації. Дане дослідження використовувало 0-1 алгоритм визначення коректності авторства публікацій, заснований на перевірці метаданих публікації за певними параметрами. Представлений в «Додатку А» на «Плакаті 4», алгоритм мав недоліки у часі роботи, складності алгоритму, кількості виконуваних операцій та звернень до бази даних, а також у точності алгоритму. Хоча точність визначення саме помилково доданих публікацій була високою, була наявна велика кількість хибних спрацьовувань, – тобто таких публікацій, що були помічені як хибні, хоча такими не були. Це пов'язано з недосконалістю методу визначення, помилках при записі інформації про публікацію у профілі Google Scholar, та відсутністю транслітерації у профілі науковців.

Дана магістерська дисертація розвиває та модифікує метод оцінки авторства публікації за допомогою введення коефіцієнту надійності для підвищення точності оцінки авторства публікацій, та зменшення кількості false-positive результатів.

Робота [12] вводить поняття графів співавторства та методи їх побудови та використання. Дана робота використовує графи співавторства для реалізації критерію 7 алгоритму Коефіцієнту надійності, який досліджує ця магістерська дисертація. Автор модифікує метод побудови матричного графу співавторства таким чином, що граф будується лише для одного наукового співробітника, а не наукового підрозділу в цілому. Також публікація представляє методи аналізу даного графу, які не використовуються в межах даної дисертації.

Одним з важливих досліджень для даної магістерської дисертації є робота [18]. Це дослідження пропонує вирішення задачі неоднозначного авторства публікацій у деяких наукометричних системах. Автор статті

пропонує використовувати наведений нижче формальний алгоритм для точної ідентифікації авторства статті.

На першому кроці виділяються елементи бібліографічного опису: імена науковців, назви публікацій, тощо. Далі відбувається пошук на предмет можливих дублікатів у системі – це дає можливість точної ідентифікації всередині системи. З бібліографічних описів виділяється ПІБ окремих авторів та будується граф співавторства. По текстовому опису відбувається пошук IRID авторів, та, за допомогою графу співавторства, виділяються найбільш імовірні пари авторів, що видаються користувачу у якості результату [18].

Дослідження [19] дає загальний опис формальних методів оцінки авторства документу, втім, для методів класифікації та кластеризації, які пропонуються в роботі, необхідно мати наявності самі тексти наукових робіт. Враховуючи велику кількість вхідних даних до даної магістерської дисертації та часту відсутність повних версій робіт у відкритому доступі, використання даних методів в рамках дослідження дисертації не є можливим.

Робота [20] робить загальний огляд можливостей використання машинного навчання та методів дослідження великих даних для аналізу наукометричних даних. Наведені моделі використовуються для загального ознайомлення з можливостями подальшого аналізу отриманих та перевірених даних.

Робота [21] містить методи досліджень авторства текстів на основі стилеметрії. Стаття пропонує метод порівняння частоти зустрічі окремих слів з списку Сводеша еталонного тексту з частотою зустрічі слів з списку Сводеша в досліджуваних текстах. Даний метод накладає важливі обмеження, а саме – досліджувані тексти мають бути одного розміру з еталонним, і метод може давати неочікувані результати на текстах з багатьма авторами. Якщо еталонний текст написаний групою авторів,

метод дозволить лише встановити, чи досліджуваний текст написаний тією самою групою [21]. Іншим недоліком є необхідність мати тексти наукових робіт, тому даний метод також не може бути використаний в рамках цієї дисертації.

Публікація [22] наводить метод визначення авторства публікації на основі інших підтверджених текстів. Автор статті розглядає лінгвометричні (лінгвометрія – галузь прикладної лінгвістики, що виявляє, вимірює та аналізує кількісні характеристики одиниць різних рівнів мови чи мовлення [22]) показники для виявлення авторської атрибуції. Це дозволяє зіставити тексти, визначивши їх авторські атрибуції, для чого вираховується частота вживання слів. На основі цих даних підтверджується авторство тексту. Хоч даний метод показує гарну точність підтвердження, втім займає багато часу, для нього необхідні повні тексти робіт, а також вручну визначені тексти робіт наукових співробітників. Даний метод гарно працює лише на роботах з одним автором, тому для цілей даної дисертації він не може бути використаний.

Робота [23] описує можливості використання графів взаємодії, які в контексті даної задачі використовуються у якості графів співавторства. Робота описує методи побудови таких графів та можливості їх використання для аналізу даних, кластеризації та будування блокових моделей для демонстрації зв'язків між науковцями.

В дослідження [24] наводиться використання мереж співавторів та визначаються основні параметри вузлів цих мереж. Дане дослідження допомагає визначити необхідні елементи перевірки для критерію 7 алгоритму оцінки коефіцієнту надійності.

Публікація [25] наводить приклад розв'язання задачі текстових атрибутів в електронних каталогах. Наведені методи можуть бути використані в даній дисертації для здійснення нечіткого пошуку та порівняння подібних рядків.

Дослідження [26] наводить приклад дослідження порівняння рядків за допомогою наближеності за N-грамами. Як заявляє автор дослідження: «Оскільки процедура визначення подібності відповіді й еталону як на основі синтаксичних, статистичних методів, так і на основі семантичних класів, фонетичних методів потребує значних обчислювальних і часових ресурсів, доцільно застосовувати швидкий, із лінійною складністю, легко масштабований алгоритм попередньої оцінки рівня подібності рядків. ... Застосування N-грам до груп символів, а не до слів, дозволяє ефективно оцінити синтаксичну подібність відповіді й еталону [26]». Тому для даної системи розглядатимуться інші рішення, спрямовані на більш швидке порівняння рядків.

Дослідження [27] наводить приклад методу для перевірки назви видавництва, з урахуванням можливих скорочень та відмінностей. Описаний метод заснований на посимвольному порівнянні рядків та перевірці кількості співпадінь. Також наводиться порівняння описаного методу і загально прийнятих – відстані Левенштейна та 3-грам. Дослідження показує низьку ефективність, малу швидкість роботи через необхідність попарної перевірки всіх рядків, низьку точність. Втім, дослідження демонструє непогані результати у використанні порівняння методом Левенштейна. Хоч метод і має деякі обмеження, а саме недостатню точність при роботі зі скороченнями та різними версіями слів, відстань Левенштейна показує гарні результати у наближеному порівнянні основної назви видавництва, без урахування номеру випуску та року.

Публікація [28] розглядає алгоритми нечіткого пошуку для оцінки подібності рядків. Для даної магістерської дисертації це є одним з важливих досліджень, так як одним з критеріїв при визначенні коефіцієнта надійності – перевірка джерела. В цілому публікація представляє можливості здійснення такого пошуку та наводить порівняння основних алгоритмів такого пошуку.

Публікація [29] описує використання відстані Левенштейна та загальні методи її використання за допомогою вбудованих методів СКБД MySQL.

Публікація [30] дає повний опис алгоритму відстані Левенштейна та наводить його математичне обґрунтування. Дане дослідження дозволяє затвердити основний алгоритм для нечіткого порівняння рядків – відстань Левенштейна.

Розглянемо публікації, що описують наступну задачу – використання транслітерації та загальні правила транслітерації. Так, публікація [31] описує проблематику транслітерації прізвищ з української мови на іноземні, що використовують латиницю. Додатково описуються основні стандарти та правила транслітерації. Так як наведені в публікації методи транслітерації є односторонніми (тобто правило транслітерації працює лише в одну сторону, через те, що різні літери українського алфавіту можуть мати один відповідник в англійському алфавіті), дана дисертація використовує узагальнений метод односторонньої транслітерації (тобто назви з англійської не транслітеруються в українську).

2.4 Опис Алгоритмів розв’язання

Для вирішення поставлених задач використовуються додаткові алгоритми, такі як алгоритми порівняння рядків з урахуванням транслітерації, алгоритм обрахунку індексу Хірша для групи науковців, алгоритм перерахунку середньої цитованості, побудови графу співавторства у матричному представлені, сортування тощо. Дані алгоритми не є основними, які досліджуються у даній роботі, а є частиною алгоритмів, що використовуються для досягнення цілей даної роботи, або ж є допоміжними алгоритмами для роботи основного алгоритму чи інформаційної системи в цілому. Розглянемо більш детально деякі з цих

алгоритмів. Для розгляду алгоритмів буде представлено формальний опис алгоритму, його призначення, графічне зображення алгоритму у вигляді блок-схеми за необхідності, код алгоритму та його приклад.

2.4.1 Перевірки на співпадіння прізвища та ініціалів

Перевірка на співпадіння прізвища та ініціалів є однією з початкових перевірок для оцінки критерію надійності публікації наукового співробітника ЗВО. Для здійснення даної перевірки необхідно порівняти записане в профілі Google Scholar ім'я наукового співробітника з інформацією про авторів з метаданих публікації. Особливість сервісу Google Scholar полягає у відсутності формалізації та стандартизації даних про публікацію та профілі, отже можливі випадки, коли ця перевірка не буде пройдена через відсутність скорочених ініціалів або повного імені тощо.

Тому необхідно при даній перевірці враховувати наступне:

- відсутність ініціалів та/або повного імені;
- дані в метаданих публікації та профілі дані записані різними мовами;
- використано різні правила транслітерації між різними мовами;
- відсутність єдиного стандарту розділових знаків для відмежування різних авторів однієї публікації.

Введемо скорочення та позначення, які будуть використані для простоти формалізованого та графічного представлення алгоритму.

Ра – Рядок авторів – метадані публікації з інформацією про авторство даної публікації.

Рп – Рядок імені профілю – ім'я або імена, представлені у описі профілю наукового співробітника ЗВО у системі Google Scholar.

Словники С1 та С2, представлені на рис. 2.1 та рис. 2.2 для транслітерації рядків Р_а та Р_п між кирилицею та латиницею.

'а' => 'a',	'б' => 'b',	'в' => 'v',
'г' => 'g',	'ґ'=>'g',	'д' => 'd', 'е' => 'e',
'є' => 'ye',	'ж' => 'zh',	'з' => 'z',
'и' => 'y',	'і'=>'i', 'ї'=>'i',	'й' => 'y', 'к' => 'k',
'л' => 'l',	'м' => 'm',	'н' => 'n',
'о' => 'o',	'п' => 'p',	'р' => 'r',
'с' => 's',	'т' => 't',	'у' => 'u',
'ф' => 'f',	'х' => 'kh',	'ц' => 'c',
'ч' => 'c',	'ш' => 's',	'щ' => 's',
'ь' => '',	'ю' => 'iu',	'я' => 'ia'

Рисунок 2.1 – Словник транслітерації 1

'а' => 'o',	'б' => 'b',	'в' => 'v',
'г' => 'h',	'ґ'=>'g',	'д' => 'd', 'е' => 'o',
'є' => 'ie',	'ж' => 'zh',	'з' => 'z',
'и' => 'y',	'і'=>'i', 'ї'=>'i',	'й' => 'y', 'к' => 'k',
'л' => 'l',	'м' => 'm',	'н' => 'n',
'о' => 'o',	'п' => 'p',	'р' => 'r',
'с' => 's',	'т' => 't',	'у' => 'u',
'ф' => 'f',	'х' => 'h',	'ц' => 'c',
'ч' => 'c',	'ш' => 's',	'щ' => 's',
'ь' => '',	'ю' => 'iu',	'я' => 'ia'

Рисунок 2.2 – Словник транслітерації 2

Дані словники засновані на офіційних правилах транслітерації України [3]. Два словника транслітерації використовується через різні правила транслітерації окремих літер в деяких іменах. Так, кирилична літера А може бути транслітерована в латиницю А або О, в залежності від Імені та початкової мови.

Формалізований опис алгоритму може бути представлений наступним чином:

- Крок 1. На першому кроці алгоритму відбувається заміна символів рядку Ра «.», «,», «/», «|», «;», «*» на пустий символ « ». Це робиться для полегшення подальшої роботи і приведення рядку авторства до одного виду.
- Крок 2. Замінюються службові символи за допомогою регулярного виразу `/[^\S\r\n]+/` на пустий символ « ». Це необхідно для того, щоб прибрати службові символи, що могли додатись при автоматизованому зборі даних про публікації.
- Крок 3. Змінений рядок Ра приводиться до нижнього регістру та обрізаються зайві пробіли на початку та в кінці.
- Крок 4. Кожен символ рядку Ра транслітерується за допомогою Словника 1, а потім Словника 2. Отримані рядки РаТ1 і РаТ2 поєднуються з початковим рядком Ра. На даному етапі отримано повністю транслітерований рядок авторів РаТ.
- Крок 5. Дії 1-4 проводяться з рядком Рп. Після заміни розділових знаків, заміни службових символів, та поєднання рядка з транслітерованим двома способами. На виході отримано повністю транслітерований рядок РпТ.
- Крок 6. На даний момент рядки РпТ та РаТ мають набір транслітерованих слів, які розділені символом пробіл « ».
- Крок 7. Розкладемо ці рядки в масиви, розбивши на слова за розділовим знаком « ».
- Крок 8. Знайдемо перетин масивів. Якщо перетин знайдено – можна зробити припущення про підтвердження авторства публікації за першою умовою; якщо ні – модифікується коефіцієнт довіри публікації.

На рис. 2.3 представлено графічний опис алгоритму у вигляді блок-схеми алгоритму.

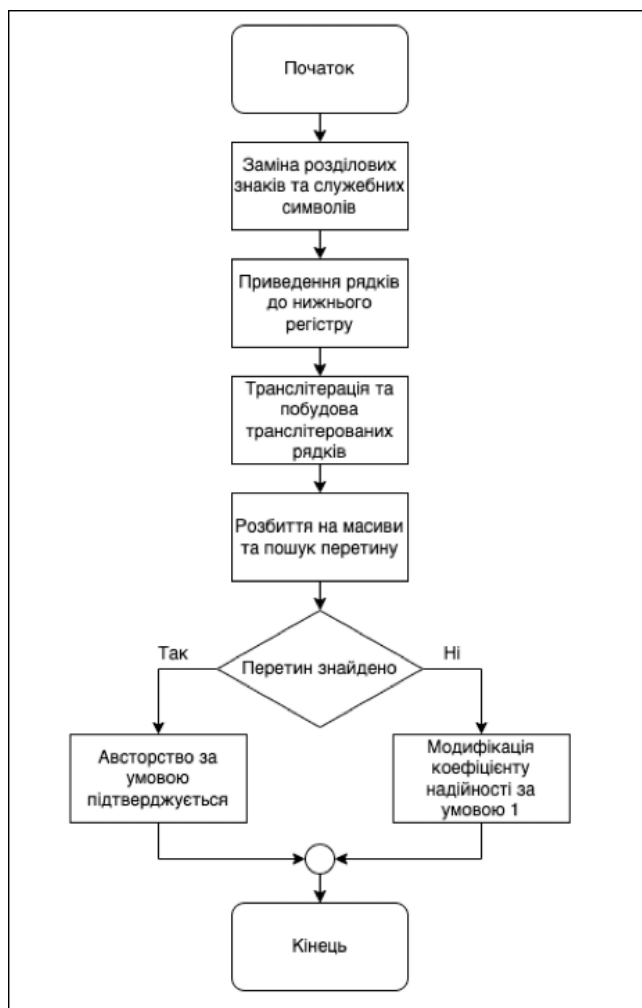


Рисунок 2.3 – Блок-схема алгоритму перевірки авторства публікації за перевіркою прізвища

Даний алгоритм, попри відносну простоту, дозволить відкинути публікації, в яких не буде співпадіння за прізвищем, або ті, що митимуть завеликі розбіжності. Втім даний алгоритм покладається на правильно заповнені дані профілю Google Scholar, що є його певним недоліком.

Для перевірки ініціалів основна відмінність полягає в формування рядку ініціалів Pi.

Крок 1. Повторюються кроки 1-4 алгоритму наведеного вище.

Крок 2. Отримані рядки розкладаються в масиви, з розділовим знаком « ».

Крок 3. Перші літери кожного елементу рядків збираються та формують з

Крок 4. Них нові рядки. Знаходиться їх перетин, такий, що послідовність літер, що перетинаються не менше 3.

Крок 5. Якщо такий перетин знайдено можна зробити висновок про

Крок 6. Підтвердження авторства публікації за другою умовою, якщо ні - модифікується коефіцієнт довіри публікації.

Графічний опис даного методу наведено у вигляді блок-схеми алгоритму на рис. 2.4.

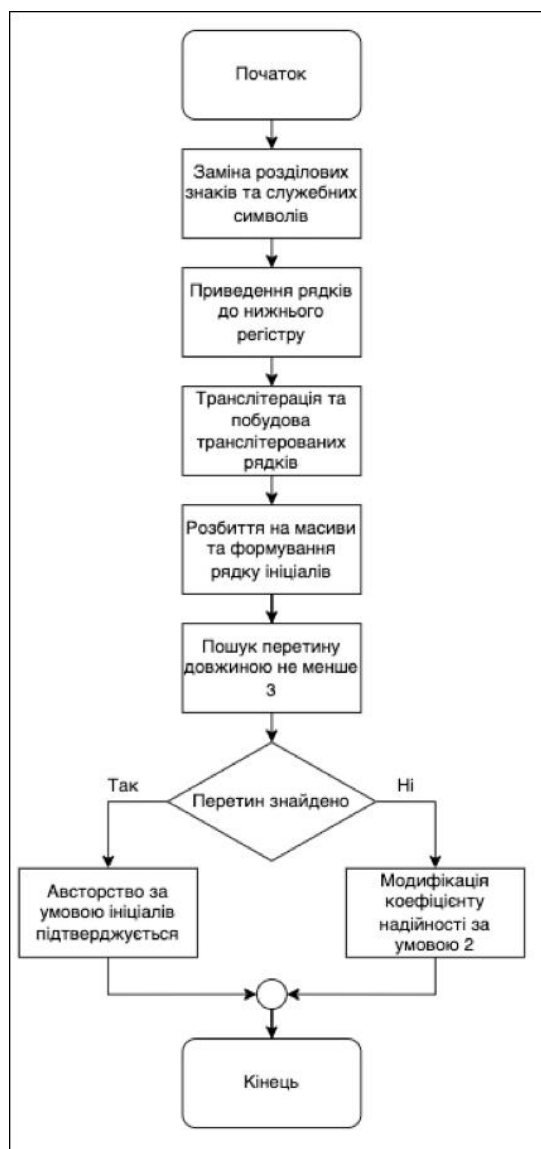


Рисунок 2.4 – Блок-схема алгоритму перевірки авторства публікації за ініціалами

Даний метод здатний відсіяти результати які не задовольняють умови перевірок – тобто визначити публікації з невірно заповненими авторами або публікації з хибним авторством.

Недоліком даного методу є те, що він опирається на інформацію з опису профілю та заповнені дані про авторство публікації, отже некоректне заповнення цих даних може призводити до false-positive результатів для даної перевірки.

Наведемо приклади даних перевірок. Оберемо декілька публікацій, які мають і які не мають пройти перевірку. На рис 2.5 зображено публікацію, взяту з системи Google Scholar.

Метод визначення відстані до об'єктів за допомогою характеристикних рис в стереозображеннях	ХК В.	Системний аналіз та інформаційні технології: 17-а міжнародна науково ..., 2015	2015	0
---	-------	--	------	---

Рисунок 2.5 – Приклад публікації з графічного інтерфейсу підсистеми

Рис. 2.6 зображає публікацію у системі Google Scholar.

Метод визначення відстані до об'єктів за допомогою характеристикних рис в стереозображеннях	2015
ХК В.	
Системний аналіз та інформаційні технології: 17-а міжнародна науково ...	

Рисунок 2.6. Приклад публікації в системі Google Scholar

В рядку авторства публікації зазначено авторів з помилкою в оформленні, таким чином метод не може однозначно оцінити та підтвердити авторство даної публікації та позначає її як таку, що не пройшла перевірку на прізвище.

В описі даного профілю вказане лише ім'я та прізвище двома мовами – Kostyantyn Kharchenko і Костянтин Харченко.

Розглянемо перевірку методом більше детально. На етапі перевірки прізвища ми отримаємо транслітерований рядок імені профілю:

kostyantyn kharchenko костянтин харченко kostyantyn kharchenko kostyntin harchenko kostyantyn kharchenko kostyntin horconko

Після розбиття даного рядку на масив імен отримаємо наступний набір імен зображений на рис 2.7.

```
0 => "kostyantyn"
1 => "kharchenko"
2 => "костянтин"
3 => "харченко"
4 => "kostyantyn"
5 => "kharchenko"
6 => "kostyntin"
7 => "harchenko"
8 => "kostyantyn"
9 => "kharchenko"
10 => "kostyntin"
11 => "horconko"
```

Рисунок 2.7 – Розбитий масив імен з транслітерованого рядка імені профілю

Розбивши імена вказані в авторстві публікації отримаємо наступний масив імен зображений на рис 2.8.

```
0 => "ХК"
1 => "В"
```

Рисунок 2.8 – Масив імен авторів публікації

Як можна побачити перетин даних масивів дасть нульовий результат, отже ми не можемо підтвердити достовірність авторства публікації за прізвищем.

Проведемо перевірку за ініціалами. Згідно алгоритму, на першому кроці ми маємо транслітерований рядок перших літер:

Kkxxkkkhhkkkhhkkkhhkkkhhkkkhhkkkhhkkkhh

Рядок імен-ініціалів аналогічний рисунку 2.9. Таким чином, перетину не знайдено і підтвердити авторство за ініціалами не вдалось. Розглянемо це більш детально.

Через помили ведення профілю, автор публікації в системі Google Scholar зазначений як «ХК В.». Тобто ім'я об'єдналось з прізвищем, і окремо було записане ім'я по батькові. Таким чином, підтвердити авторство за ініціалами не вдалось.

Наведемо приклад публікації, що проходить перевірку на прізвище, але не проходить перевірку на ініціали. Прикладом такої публікації буде публікація, зображена на рис. 2.9.

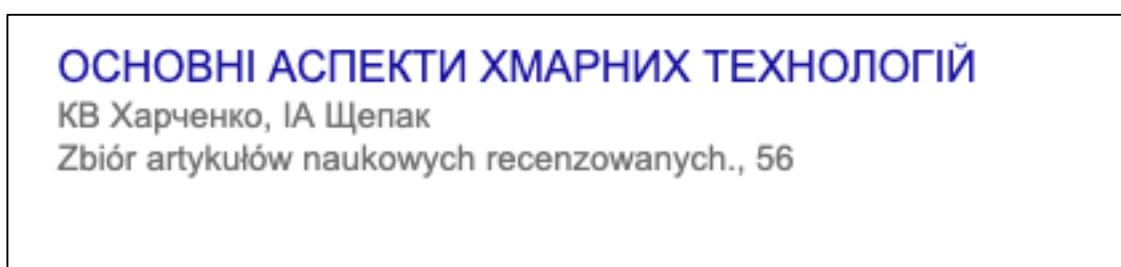


Рисунок 2.9 – Публікація, яка не проходить перевірку на ініціали

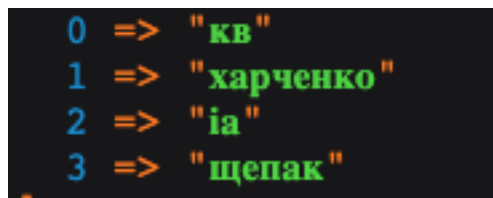
Транслітерований рядок профілю виглядає наступним чином:

*kostyantyn kharchenko косянтин харченко kostyantyn kharchenko kostyntin
harcenko kostyantyn kharchenko kostyntin horconko*

Транслітерований рядок ініціалів має наступний вигляд:

Kkxxkkkhhkkkhhkkkhhkkkhhkkkhhkkkhhkkkhh

Список імен авторів публікації представлений на рис. 2.10.



```
0 => "кв"
1 => "харченко"
2 => "іа"
3 => "щепак"
```

Рисунок 2.10 – Автори публікації, представленої на рисунку 2.9

Можна побачити, що є перетин по прізвищу, але відсутній перетин по ініціалах. Перетин по ініціалах не було знайдено, тому що в авторах публікації вказані ініціали для імені та імені по-батькові, а в описі профілю ім'я по-батькові відсутнє, тому не можна підтвердити публікацію за ініціалами.

Нарешті розглянемо варіант коли публікація пройде обидві перевірки. Приклад такої публікації наведено на рис 2.11.

A set of instructions for data flow virtual machine

5

2017

K Kharchenko, O Beznosyk, V Romanov

2017 IEEE First Ukraine Conference on Electrical and Computer Engineering ...

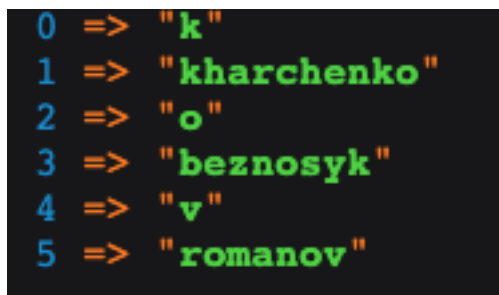
Рисунок 2.11– Публікація, яка проходить обидві перевірки – на прізвище та ініціали

Наведемо транслітеровані рядки профілю та ініціалів

*kostyantyn kharchenko косянтин харченко kostyantyn kharchenko kostyntin
harcenko kostyantyn kharchenko kostyntin horconko*

kkkxkkkhkkkhkkkhkkkhkkkhkkkhkkkhkkkh

Список авторів представлено на рис. 2.12.



```
0 => "k"
1 => "kharchenko"
2 => "o"
3 => "beznosyk"
4 => "v"
5 => "romanov"
```

Рисунок 2.12 – Список авторів публікації, представленої на рис. 2.11

Як можна побачити, для даної публікації буде перетин і з рядком імені профілю і з рядком ініціалів. Тобто дана публікація пройде обидві перевірки і алгоритм може підтвердити авторство публікації на основі порівняння імені з ім'ям зазначеним у профілі наукових співробітників.

2.4.2 Перевірка джерела публікації

Однією з перевірок у алгоритмі оцінки надійності публікації є перевірка джерела публікації на наявності інших робіт даного автора у ньому.

Так, наявність інших публікацій в джерелі дозволить зробити припущення про потенційно суміжну сферу знань, на тему якої написана дана публікація та те, що науковий співробітник ЗВО вже публікувався в цьому джерелі, тож потенційно ця публікація є коректною.

Метадані джерела публікації, зазвичай, представлені у наступному форматі: Назва журналу, випуск, сторінки (за наявності), рік.

Враховуючи, що необхідно підтвердити лише назву журналу, достатньо визначити подібність рядків метаданих публікації, що перевіряється, та інших публікацій наукового співробітника ЗВО.

Оптимальним методом для пошуку подібності рядків такого роду є пошук відстані Левенштейна, або редакторської відстані, між рядками метаданих джерела публікації. Оскільки необхідно порівняти лише назву джерела без урахування номеру випуску, року, сторінок тощо, зробимо припущення, що для підтвердження коректності порівняння двох рядків джерел публікацій, вони мають співпадати не менше, ніж на 70%.

Введемо наступні умовні позначення:

- R_s - Рядок джерела публікації, що розглядається;
- *levenshtein* - відстань Левенштейна для джерела публікації, що розглядається і публікації з якою відбувається порівняння.

Формалізовано алгоритм перевірки виглядає наступним чином.

Крок 1. Для усіх публікацій Автора зібрати метадані про джерело публікації в окремий масив рядків.

Крок 2. Порівняти рядки метаданих джерела публікації для публікації, що перевіряється з кожним рядком в зібраному масиві метаданих джерела публікацій наукового співробітника ЗВО.

Крок 3. Розрахувати відстань Левенштейна для рядків метаданих джерела публікації, для якої здійснюється перевірка з зібраними метаданими джерела інших публікацій.

Крок 4. Розрахувати відношення відстані Левенштейна до довжини рядка за формулою 3.

$$d = \frac{\textit{levenshtein} * 100}{\textit{len}(R_s)} \quad (3)$$

Крок 5. Якщо відношення d буде менше 30 можна зробити висновок, що дві публікації в одному джерелі.

Крок 6. Якщо знайдена хоча б одна публікація, окрім тієї, для якої робиться перевірка, можна зробити висновок, що у наукового

співробітника ЗВО є декілька публікацій в даному джерелі, а отже публікація є підтвердженою за даним критерієм.

Представимо описаний вище алгоритм у графічному вигляді за допомогою блок-схеми алгоритму на рис. 2.13.

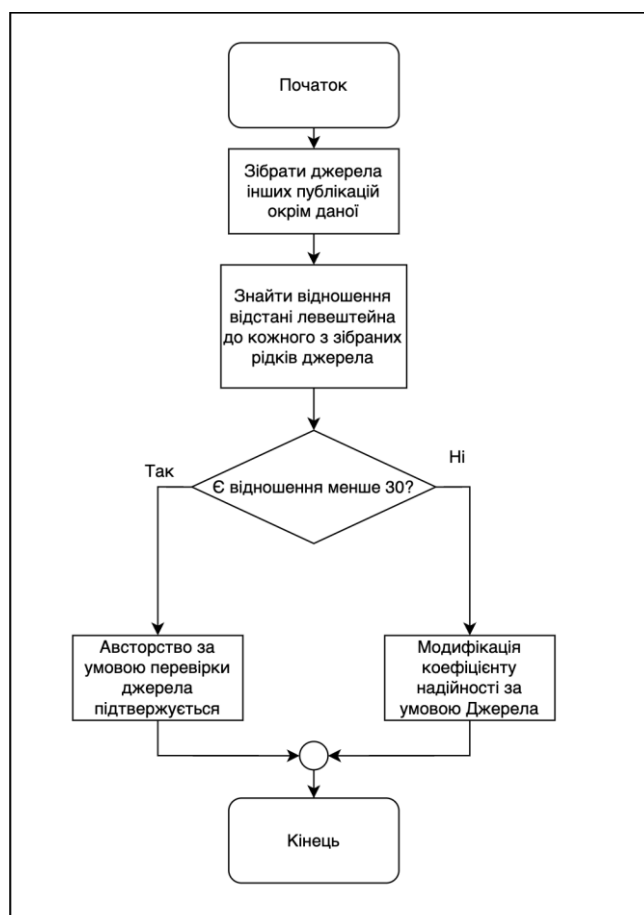


Рисунок 2.13 – Блок-схема алгоритму перевірки джерела публікації

Наведемо нижче приклад перевірок за критерієм джерела публікації. Перевіримо публікацію науковця на ім'я Колосов Олександр Євгенович, дані про яку наведені на рис. 2.14

Виброакустический эффект при ультразвуковой пропитке волокнистых композитов	АЕ Колосов, ВГ Хозин, АА Каримов	Механика композиционных материалов, 743-746, 1987	1987	4
---	--	---	------	---

Рисунок 2.14 – Приклад публікації Колосова О.Є.

Дана публікація має джерело «Механика композиционных материалов, 743-746, 1987», надалі джерело 1.

Провівши перевірку даним алгоритмом, можна побачити що у даного наукового співробітника є ще дві публікації в даному джерелі. Наведемо їх на рис. 2.15 та рис. 2.16.

Некоторые аспекты	АЕ Колосов, ВВ	Механика композит.	1988	5
детерминирования адекватной	Клявлин	материалов, 1012-1019,		
модели структуры		1988		
ориентированных волокнистых				
композитов				

Рисунок 2.15 – Перша знайдена публікація з джерелом 1

Пропитка волокнистых	АЕ Колосов, ВГ	Механика	1988	19
наполнителей полимерными	Хозин, АА	композиционных		
связующими	Каримов, ВВ	материалов, 651-659,		
	Клявкин	1988		

Рисунок 2.16 – Друга знайдена публікація з спільним джерелом 1

Як можна побачити, в даних публікаціях відрізняються сторінки та рік публікації. А в початковій публікації та першій знайдений є незначна відмінність у записі джерела, в ньому присутнє скорочення.

В першому випадку відстань Левенштейна буде дорівнювати 22, а відношення 27.16, що дозволяє нам вважати що у даних публікацій одне джерело.

В другому випадку відстань Левенштейна рівна 7, а відношення 8.64, що також дає можливість зробити припущення про спільне джерело в даних публікаціях.

Розглянемо приклад публікації на рис. 2.17, яка не пройшла дану перевірку, щоб переконатись в вірності визначених критеріїв та правильності методу.

Методика расчета ультразвукового кавитационного устройства с излучающей пластиной	АФ Луговской, АЕ Колосов	Экотехнологии и ресурсосбережение, 59-67, 2005	2005	3
---	-----------------------------	--	------	---

Рисунок 2.17 – Приклад публікації з відмінним джерелом від джерела 1

Відстань Левенштейна джерел з даною публікацією буде дорівнювати 52, а відношення 64.19.

Отже можна зробити висновок про коректне визначення спільності джерела публікації.

2.4.3 Детальний опис пошуку індексу Хірша для підрозділу

Для аналізу даних по групам наукових підрозділів (наприклад кафедр або факультетів) необхідно мати можливість розрахувати індекс Хірша не лише для одного наукового співробітника, а всього підрозділу ЗВО. Вище приводилась математична модель розрахунку індексу Хірша для групи підрозділу. Нижче наведено більш детальний алгоритм розрахунку індексу Хірша для наукового підрозділу або групи наукових співробітників ЗВО.

Формально прикладний алгоритм оцінки індексу Хірша для наукового підрозділу можна описати наступним чином:

Крок 1. Зібрати інформацію про унікальні публікації наукових співробітників певної наукової групи. Під унікальною публікацією мається на увазі, що публікації за авторством двох

наукових співробітників, що входять в наукову групу, необхідно врахувати лише один раз.

Крок 2. Відсортувати зібрані публікації за кількість цитувань.

Крок 3. Знайти таку публікацію з найвищим порядковим номером h , яка матиме не менше h цитувань.

Крок 4. Число h і буде показником індексу Хірша для даного підрозділу або групи наукових співробітників ЗВО.

Представимо даний алгоритм у графічному вигляді за допомогою блок-схеми на рис. 2.18.

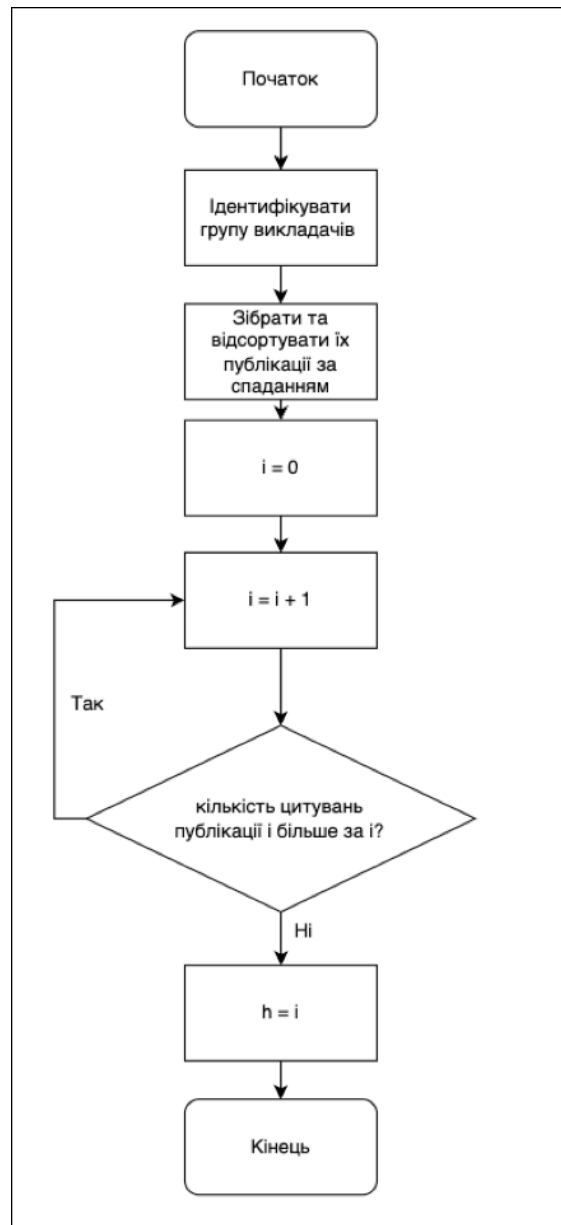


Рисунок 2.18 – Блок-схема алгоритму обрахунку індексу Хірша для групи авторів

Для простоти розуміння даного алгоритму нижче наведено його псевдокод.

```

count = count(publications);
h = count+1;
change=0;
while (!change) {
    current_link = 0;
    foreach (publications as pub) {

```

```

    if (pub['citations'] > current_link) {
        current_link += 1;
    }
    if (current_link < h) {
        h -= 1;
    } else {
        change=1;
    }
}
}
}

```

В результаті ми отримано h - індекс Хірша для групи авторів, які мають список публікацій publication.

Наведемо приклад пошуку індексу Хірша для групи наукових співробітників ЗВО. Оберемо у якості групи наукових співробітників кафедру «загальної та неорганічної хімії ХТФ». Алгоритм показує, що дана кафедра матиме індекс Хірша, рівний 7.

На рис. 2.19 можна побачити показники цієї кафедри та інших кафедр факультету.

Google Scholar рейтинг кафедр на факультеті xtf			
Кафедра	статей	цитування	h-індекс
технології електрохімічних виробництв ХТФ	672	1127	12
фізичної хімії ХТФ	577	1409	13
хімічної технології композиційних матеріалів ХТФ	591	1194	13
загальної та неорганічної хімії ХТФ	326	335	7
органічної хімії та технології органічних речовин ХТФ	504	8364	31
технології неорганічних речовин, водорочищення та загальної хімічної технології ХТФ	1079	2361	15
хімічної технології кераміки та скла ХТФ	390	1694	15
Хіміко-технологічний факультет	0	0	0
Хіміко-технологічний факультет // Сертифікована науково-дослідна лабораторія адсорбції та іонного обміну ХТФ	0	0	0

Рисунок 2.19 – Кафедри факультету ХТФ

На рис. 2.20 зображені публікації наукових співробітників цієї кафедри.

name	citations
1 New carbon based materials for electrochemical energy storage systems: batteries, superc...	71
2 Hydrogen plasma treatment of silicon thin-film structures and nanostructured layers	19
3 Thermodynamic properties of multiwalled carbon nanotubes	13
4 The study of low-temperature heat capacity of diamond: Calculation and experiment	12
5 Formation of nano-sized oxides in K-Ta-O system by chemical reaction of Ta metal with KN...	11
6 Crystal structure of 9, 9'-{(1E, 1' E)-[1, 4-phenylenebis (azanylylidene)] bis (methanyl...	11
7 Formation of spinel structured compounds in the lithium permanganate thermal decompositi...	8
8 Синие-зеленые неорганические пигменты, синтезированные с использованием отходов гальванич...	7
9 Спектроскопические исследования пф̄-аминоэтилатных комплексов кобальта(III)	7
10 Special features of the heat capacity of detonation nanocrystalline diamond	6
11 Катализаторы электровосстановления молекулярного кислорода на основе гетеробиметаллическ...	6
12 Стрoение внутрикoмплексного соединения кобальта(III) с триэтанoламином	6
13 Discharge-ionization of hydrogen on modified carbon nanotube electrodes	6
14 Preparation and properties of high-dispersive powders of aluminium dodecaoride and carbo...	5
15 Синтез та фізико-хімічні властивості нанодисперсних оксидів титану, стануму, танталу	5

Рисунок 2.20 – Публікації кафедри Загальної та неорганічної хімії

На рис. 2.20 можна побачити, що публікація з порядковим номером 7 має 8 цитувань, наступна ж публікація під номером 8 має 7 цитувань. Необхідно обрати публікацію з номером h таку, що кількість цитувань цієї публікації буде не менше за h . Для даного списку публікацій число h буде рівне 7. Таким чином, індекс Хірша для даного підрозділу буде рівний 7.

Наведений приклад наочно демонструє розрахунок індексу Хірша для наукового підрозділу або групи наукових співробітників ЗВО. Використання даного методу дозволить аналізувати факультети, кафедри, окремі наукові підрозділи або цілі університети за показником індексу Хірша.

2.4.4 Побудова графу співавторства у матричному вигляді

Як було наведено вище, граф співавторства – графова структура даних, яка показує зв'язки наукових співробітників ЗВО між собою. Наукові співробітники представлені вершинами графу, а їх спільні публікації – їх

зваженими зв'язками, де вагою ребра даного графу буде кількість спільних публікацій між парою наукових співробітників ЗВО.

Наведемо формальний алгоритм побудови такого графу у матричному представленні. Метод валідації використовує графи співавторства на основі двох рукостискань – тобто такі, які представляють наукового співробітника, його співавторів і їх співавторів.

На першому кроці необхідно зібрати всі публікації наукового співробітника та знаходимо його співавторів.

За допомогою унікальних ідентифікаторів наукового співробітника побудуємо та ініціалізуємо наступну матрицю.

$$\begin{array}{ccccc}
 - & a_1 & a_2 & \cdots & a_n \\
 a_1 & 0 & 0 & \cdots & 0 \\
 a_2 & 0 & 0 & \cdots & 0 \\
 \cdots & \cdots & \cdots & \cdots & \cdots \\
 a_n & 0 & 0 & \cdots & 0
 \end{array},$$

де a_m – унікальний ідентифікатор наукового співробітника m ;

Таким чином елементом елемент матриці a_{ij} буде показувати кількість спільних публікацій у наукових співробітників i та j .

На другому кроці необхідно пройти по всіх публікаціях наукового співробітника, та для кожної такої публікації збільшити число в відповідній клітинці a_{ij} .

Третім кроком алгоритму буде повторити дії 1-3 для кожного наукового співробітника $a_1 - a_n$.

Таким чином буде побудовано матричну форму графу співавторства для двох рукостискань.

В результаті буде отримана матриця наступного виду:

$$\begin{array}{cccccc}
 a_{11} & a_{12} & \dots & a_{1n} & \dots & a_{1m} \\
 a_{21} & a_{22} & \dots & a_{2n} & \dots & a_{2m} \\
 \dots & \dots & \dots & \dots & \dots & \dots \\
 a_{n1} & a_{n2} & \dots & a_{nn} & \dots & a_{nm} \\
 \dots & \dots & \dots & \dots & \dots & \dots \\
 a_{m1} & a_{m2} & \dots & a_{mn} & \dots & a_{mt}
 \end{array}$$

Наведемо графічне представлення у вигляді блок-схеми алгоритму на рис. 2.21.

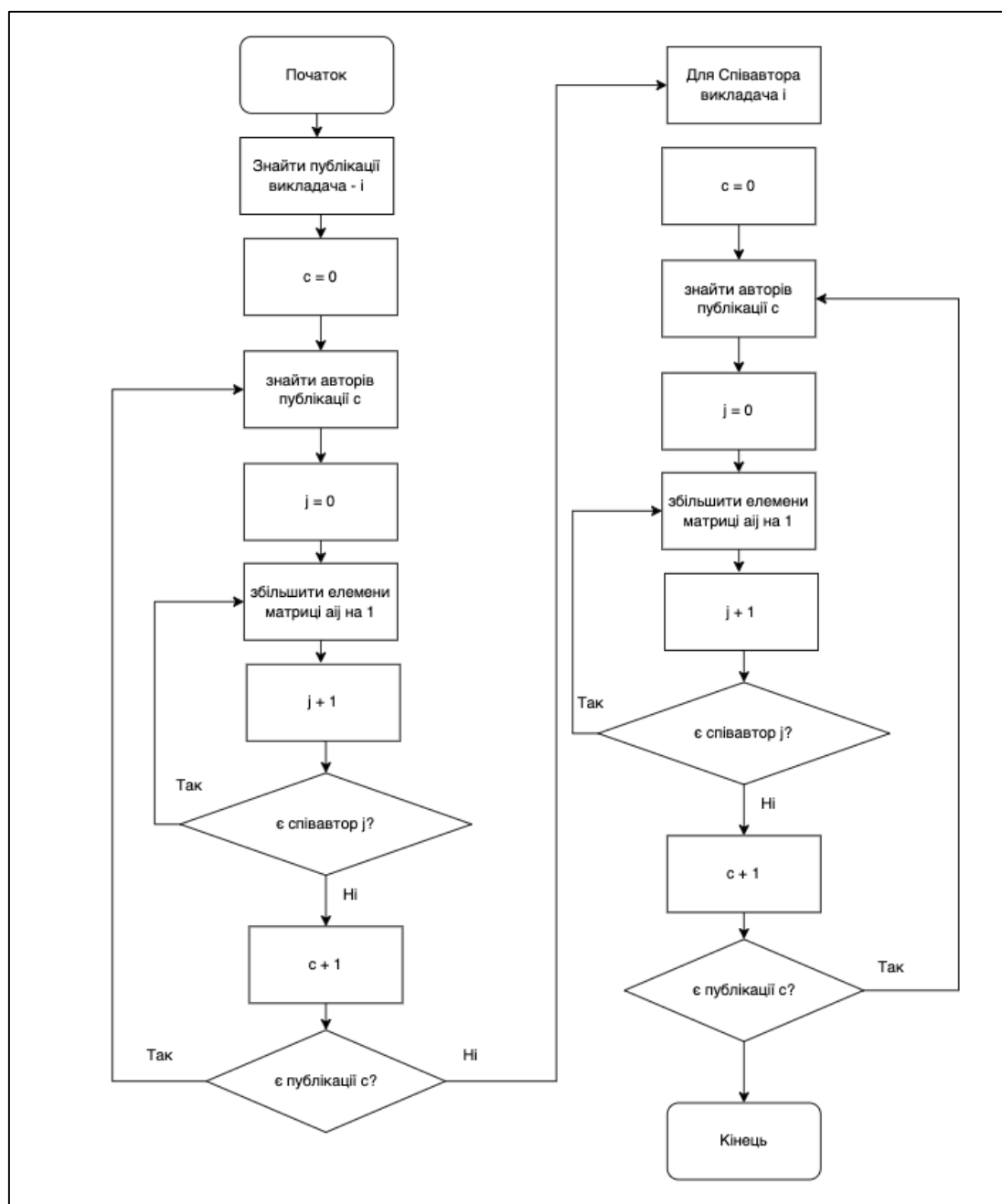


Рисунок 2.21 – Блок-схема алгоритму

Нижче наведено псевдокод даного алгоритму.

```

professor;
publications;
matrix;

for publication in publications:
    for author in publication.authors:
        if professor.id != author.id:
            matrix[professor.id][author.id] += 1

for key in matrix[professor.id]:
    key_professor;
    key_publications;
    for publication in key_publications:
        for author in publication.authors:
            if key_professor.id != author.id:
                matrix[key_professor.id][author.id] += 1

```

Таким чином, буде отримано матричне представлення графу співавторства для наукового співробітника `professor` відповідно до його публікацій.

2.5 Удосконалення методу оцінки надійності

Попередні дослідження [7] представляли алгоритм оцінки авторства публікацій на основі формалізованих параметрів, які мали на меті визначити надійність такої публікації. Основним недоліком даного алгоритму була велика кількість *false-positive* результатів. Тобто велика кількість публікацій авторства наукового співробітника позначались як ненадійні через доволі простий алгоритм, помилки в веденні профілю, помилки в метаданих публікацій тощо. Також недоліком була типізація оцінки надійності публікації – було лише два результати 0 та 1, що представляють потенційно ненадійну публікацію, та публікацію, позначену як коректну. Додатково, алгоритм не враховував випадків повних однофамільців з співпадінням прізвища, імені та імені по-батькові.

Дана робота вводить коефіцієнт надійності, описаний в пункті 2.2.2 цієї роботи.

Метод Коефіцієнту надійності публікації має більшу кількість критеріїв, а головне – вводить коефіцієнт, якій коливається в межах, які визначаються вагою критерія перевірки.

Критерії та їх вага визначені в таблиці 2.1. Варто зазначити, що вказана вага критеріїв є початковою. За допомогою наступних досліджень проведемо налаштування параметрів критеріїв алгоритму для того, щоб вони були оптимальними.

Попередні дослідження та наявні математичні моделі, в особливості алгоритм 2.4.3, демонструє, що перевірка за критерієм 6 не має сенсу, оскільки виключення однієї публікації з списку для побудови індексу Хірша, у більшості випадків, не матиме на нього впливу. Єдиний випадок, коли виключення публікації може змінити індекс Хірша – це виключення публікації під номером h з загального списку дослідження h -індексу.

Це може призвести до зміни індексу Хірша у рамках $[h-1; h+1]$. Згідно з критерієм, необхідно оцінити публікацію, вилучення якої змінить індекс Хірша більше, ніж на 50%. Якщо враховувати можливість зміни показника індексу Хірша лише в діапазоні $[h-1; h+1]$, даний критерій не буде пройдений лише у випадку, якщо індекс Хірша рівний 2.

Таким чином, дана перевірка не матиме сенсу, і може бути вилучена.

Також звернімо увагу на занадто великий вплив критерію 7. Наявність великої кількості співавторів загалом може підвищити показник довіри до публікації, втім відсутність достатньої кількості співавторів у наукового співробітника загалом не дає змоги однозначно стверджувати про наявність проблем з публікацією. Тому залишимо даний критерій на публікацію, але зменшимо його вплив. Визначимо максимальне значення цього коефіцієнта як 1, для 10 співавторів – тобто вплив на коефіцієнт надійності відсутній у випадку наявності 10 суміжних ребер на графі спів та мінімальне значення

0.9 у випадку відсутності визначених співавторів у наукового співробітника взагалі. Таким чином, виведемо новий коефіцієнт за допомогою формули 4.

$$k_7 = ((0.01 * n) + 0.9) * x \quad (4)$$

,де n – кількість ребер у графа співавторства у даного наукового співробітника з іншими науковими співробітниками;

x – поточний показник коефіцієнту надійності.

Провівши перевірки, отримані результати показують велику кількість false-positive результатів – тобто публікацій з низьким рівнем коефіцієнту довіри, без порушення важливих критеріїв.

Так, відсутність співавтора у ЗВО, не співпадіння ініціалів та одна публікація в джерелі призводила до зниження коефіцієнту довіри до 0.3456, а порушення критерію 5 та наявність лише однієї публікації в джерелі – до 0.5184. Порушення критеріїв 3, 2 та 5 – до показника в 0.5184, а порушення лише критерію 4 – до 0.6825.

Також необхідно враховувати вплив критерію 6, що впливає на усі публікації наукового співробітника.

Дані публікації продемонстровані на рис. 2.22 та рис. 2.23.

Назва	Автори	Джерело	Рік	Цитат	Індекс надійності	Коментар
Dual (triple) integral equations	NA Virchenko	Vyshcha Shkola, Kiev, 1989	1989	13	0.3456	[Неспівпадіння ініціалів] [відсутність співавтора в КПІ][одна публікація в джерелі або джерело не розпізнане]
Some integral operators of Buschman-Erdélyi type	N Virchenko	REVISTA TECNICA-FACULTAD DE INGENIERIA UNIVERSIDAD DEL ZULIA 19, 85-90, 1996	1996	1	0.5184	[відсутність співавтора в КПІ][різниця між середньою цитованістю та данною публікацією більше 50%][одна публікація в джерелі або джерело не розпізнане]
PROPERTIES OF THE GENERALIZED MELER-FOCK INTEGRAL TRANSFORMATION AND ITS APPLICATION TO THE SOLUTION OF DUAL INTEGRAL-EQUATIONS	NA Virchenko, SP Ponomarenko	DOPOVIDI AKADEMII NAUK UKRAINSKOI RSR SERIYA A-FIZIKO-MATEMATICHNI TA ..., 1979	1979	2	0.41472	[Неспівпадіння ініціалів] [відсутність співавтора в КПІ][різниця між середньою цитованістю та данною публікацією більше 50%]
ON SOME ASSOCIATED LEGENDRE-WRIGHT FUNCTIONS	N Virchenko	Издательство "УНИПРЕСС", 2000	2000	0	0.576	[відсутність співавтора в КПІ][одна публікація в джерелі або джерело не розпізнане]

Рисунок 2.22 – Результати першого етапу перевірок

Назва	Автори	Джерело	Рік	Цитат	Індекс надійності	Коментар
Tests for m-accretive closedness of a second-order linear elliptic operator	VF Kovalenko, YA Semenov	Siberian Mathematical Journal 31 (2), 249-260, 1990	1990	0	0.91	
ПРИЗНАКИ то-АККРЕТИВНОЙ ЗАМЫКАЕМОСТИ ЛИНЕЙНОГО ЭЛЛИПТИЧЕСКОГО ОПЕРАТОРА ВТОРОГО ПОРЯДКА	ВФ КОВАЛЕНКО, ЮА СЕМЕНОВ	Сибирский математический журнал 31, 76, 1990	1990	0	0.6825	[одна публікація в джерелі або джерело не розпізнане]
C0-полугруппы в пространствах $L^p(\mathbb{R}^d)$ и $\widehat{C}(\mathbb{R}^d)$, порождаемые дифференциальным выражением $\Delta + b \nabla \cdot \dots \nabla \in \mathbb{R}^d$	ВФ Коваленко, ЮА Семенов	Теория вероятностей и ее применения 35 (3), 449-458, 1990	1990	0	0.6825	[одна публікація в джерелі або джерело не розпізнане]
On LP-theory of Schrodinger semigroups	VF Kovalenko, YA Semenov	I, Ukrain. Math. J 41, 273-278, 1989	1989	5	0.61425	[різниця між середньою цитованістю та данною публікацією більше 50%] [одна публікація в джерелі або джерело не розпізнане]

Рисунок 2.23 – Результати першого прогону перевірок на іншому профілі

Отже, можна зробити висновок про необхідність зменшення впливу певних критеріїв.

У зв'язку з помилками ведення профілів, зменшимо вплив критерію ініціалів та відсутності співавтора в ЗВО. Встановимо дані критерії 0.75 та 0.85 відповідно.

Після цього публікація що порушує лише критерій 3 матиме коефіцієнт довіри на рівні 0.77 (з урахуванням модифікатора критерію 6). Втім, порушення критеріїв 3 та 4 призводитиме до визначення коефіцієнту довіри на рівні 0.58, що не є виправданим.

Тому зменшимо вплив критерію 3 до рівня 0.9, а критерію 4 до рівня 0.85.

Приведемо проміжні результати на рис. 2.24 та рис. 2.25.

Назва	Автори	Джерело	Рік	Цитат	Індекс надійності	Коментар
Tests for m-accretive closedness of a second-order linear elliptic operator	VF Kovalenko, YA Semenov	Siberian Mathematical Journal 31 (2), 249-260, 1990	1990	0	0.819	[відсутність свіпавтора в КП]
ПРИЗНАКИ то-АККРЕТИВНОЙ ЗАМЫКАЕМОСТИ ЛИНЕЙНОГО ЭЛЛИПТИЧЕСКОГО ОПЕРАТОРА ВТОРОГО ПОРЯДКА	ВФ КОВАЛЕНКО, ЮА СЕМЕНОВ	Сибирский математический журнал 31, 76, 1990	1990	0	0.69615	[відсутність свіпавтора в КП] [одна публікація в джерелі або джерело не розпізнане]
СО-полугруппы в пространствах $L^p(\mathbb{R}^d)$ и $\widehat{C}(\mathbb{R}^d)$, порождаемые дифференциальным выражением $O'' + b \nabla \cdot \dots \nabla \in$	ВФ Коваленко, ЮА Семенов	Теория вероятностей и ее применения 35 (3), 449-458, 1990	1990	0	0.69615	[відсутність свіпавтора в КП] [одна публікація в джерелі або джерело не розпізнане]
On LP-theory of Schrodinger semigroups	VF Kovalenko, YA Semenov	I, Ukrain. Math. J 41, 273-278, 1989	1989	5	0.626535	[відсутність свіпавтора в КП] [різниця між середньою цитованістю та данною публікацією більше 50%][одна публікація в джерелі або джерело не розпізнане]

Рисунок 2.24 – Вплив змінених критеріїв алгоритму, приклад 1

Назва	Автори	Джерело	Рік	Цитат	Індекс надійності	Коментар
Dual (triple) integral equations	NA Virchenko	Vyshcha Shkola, Kiev, 1989	1989	13	0.5508	[Неспівпадіння ініціалів] [відсутність співатора в КП][одна публікація в джерелі або джерело не розпізнане]
Some integral operators of Buschman-Erdélyi type	N Virchenko	REVISTA TECNICA-FACULTAD DE INGENIERIA UNIVERSIDAD DEL ZULIA 19, 85-90, 1996	1996	1	0.66096	[відсутність співатора в КП][різниця між середньою цитованістю та данною публікацією більше 50%][одна публікація в джерелі або джерело не розпізнане]
PROPERTIES OF THE GENERALIZED MELER-FOCK INTEGRAL TRANSFORMATION AND ITS APPLICATION TO THE SOLUTION OF DUAL INTEGRAL-EQUATIONS	NA Virchenko, SP Ponomarenko	DOPOVIDI AKADEMII NAUK UKRAINSKOI RSR SERIYA A-FIZIKO-MATEMATICHNI TA ..., 1979	1979	2	0.5832	[Неспівпадіння ініціалів] [відсутність співатора в КП][різниця між середньою цитованістю та данною публікацією більше 50%]
ON SOME ASSOCIATED LEGENDRE-WRIGHT FUNCTIONS	N Virchenko	Издательство" УНИПРЕСС", 2000	2000	0	0.7344	[відсутність співатора в КП][одна публікація в джерелі або джерело не розпізнане]

Рисунок 2.25 – Вплив змінених критеріїв алгоритму, приклад 2

Розглянемо детальніше показних середньої цитованості. Він є виправданим, бо дозволяє помітити публікації, що мають сильний вплив на загальне та середню цитованість наукового співробітника. Наявність такої публікації може бути коректною, але в купі з порушеннями ряду інших критеріїв має помітити публікацію, як потенційно не надійну. Така перевірка виконується для оцінки сфери знань на рівні з перевіркою джерела видавництва публікації. Так, прикладні сфери знань як правило мають менший загальний рівень цитованості за гуманітарні сфери, а вузькопрофільні роботи – ще менший. Втім, необхідно зауважити, що це є суто припущенням, тому вплив даної перевірки має бути зменшений до рівня 0.95. Також це показує, що вплив джерела публікації може бути збільшений та встановлений на рівні 0.8, так як журнал публікації може більш явно показати наукову сферу і наявність інших публікацій в цьому джерелі можуть свідчити про наявні роботи в даній сфері.

Результати перевірки зображено на рис. 2.26 та рис. 2.27 відповідно.

Назва	Автори	Джерело	Рік	Цитат	Індекс надійності	Коментар
Tests for m-accretive closedness of a second-order linear elliptic operator	VF Kovalenko, YA Semenov	Siberian Mathematical Journal 31 (2), 249-260, 1990	1990	0	0.8645	[відсутність свіпавтора в КП]
ПРИЗНАКИ то-АККРЕТИВНОЙ ЗАМЫКАЕМОСТИ ЛИНЕЙНОГО ЭЛЛИПТИЧЕСКОГО ОПЕРАТОРА ВТОРОГО ПОРЯДКА	ВФ КОВАЛЕНКО, ЮА СЕМЕНОВ	Сибирский математический журнал 31, 76, 1990	1990	0	0.6916	[відсутність свіпавтора в КП] [одна публікація в джерелі або джерело не розпізнане]
C0-полугруппы в пространствах $L^p(\mathbb{R}^d)$ и $\widehat{C}(\mathbb{R}^d)$, порождаемые дифференциальным выражением $O''+b\nabla\ldots\nabla\in^*$	ВФ Коваленко, ЮА Семенов	Теория вероятностей и ее применения 35 (3), 449-458, 1990	1990	0	0.6916	[відсутність свіпавтора в КП] [одна публікація в джерелі або джерело не розпізнане]
On LP-theory of Schrodinger semigroups	VF Kovalenko, YA Semenov	I, Ukrain. Math. J 41, 273-278, 1989	1989	5	0.62244	[відсутність свіпавтора в КП] [різниця між середньою цитованістю та данною публікацією більше 50%][одна публікація в джерелі або джерело не розпізнане]

Рисунок 2.26 – Вплив змінених критеріїв алгоритму, приклад 1

Назва	Автори	Джерело	Рік	Цитат	Індекс надійності	Коментар
Dual (triple) integral equations	NA Virchenko	Vyshcha Shkola, Kiev, 1989	1989	13	0.5472	[Неспівпадіння ініціалів] [відсутність свіпавтора в КП][одна публікація в джерелі або джерело не розпізнане]
Some integral operators of Buschman-Erdélyi type	N Virchenko	REVISTA TECNICA-FACULTAD DE INGENIERIA UNIVERSIDAD DEL ZULIA 19, 85-90, 1996	1996	1	0.65664	[відсутність свіпавтора в КП][різниця між середньою цитованістю та данною публікацією більше 50%][одна публікація в джерелі або джерело не розпізнане]
PROPERTIES OF THE GENERALIZED MELER-FOCK INTEGRAL TRANSFORMATION AND ITS APPLICATION TO THE SOLUTION OF DUAL INTEGRAL-EQUATIONS	NA Virchenko, SP Ponomarenko	DOPOVIDI AKADEMII NAUK UKRAINSKOI RSR SERIYA A-FIZIKO-MATEMATICHNI TA ..., 1979	1979	2	0.6156	[Неспівпадіння ініціалів] [відсутність свіпавтора в КП][різниця між середньою цитованістю та данною публікацією більше 50%]
ON SOME ASSOCIATED LEGENDRE-WRIGHT FUNCTIONS	N Virchenko	Издательство" УНИПРЕСС", 2000	2000	0	0.7296	[відсутність свіпавтора в КП][одна публікація в джерелі або джерело не розпізнане]

Рисунок 2.27 – Вплив змінених критеріїв алгоритму, приклад 2

Рис. 2.28 показує приклад профілю з великою кількістю успішно пройдених перевірок.

Назва	Автори	Джерело	Рік	Цитат	Індекс надійності	Коментар
ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ЯК НЕОБХІДНИЙ ІНСТРУМЕНТ ДЛЯ ЕФЕКТИВНОГО ВИКОРИСТАННЯ БАЗ ДАНИХ СИСТЕМ SCADA	ЛМ Бугаєва, ЮО Безносик	ВВК 91, 422, 2019	2019	0	0.99	
Можливості нечіткої логіки для рішення технологічних задач на прикладі газотурбінного устаткування	ЛМ Бугаєва, ЮО Безносик, МВ Іванов, ЛН Бугаєва, ЮА Безносик, ...	КПІ ім. Ігоря Сікорського, 2019	2019	0	0.99	
Комп'ютерне моделювання виробничого процесу отримання стирену.	ДП Крамаренко, ЛМ Бугаєва, ЮО Безносик	Комп'ютерне моделювання в хімії і технологіях та системах сталого розвитку ..., 2020	2020	0	0.9405	[відсутність співавтора в КПІ]
КОМП'ЮТЕРНЕ МОДЕЛЮВАННЯ ПРОЦЕСУ ОКИСЛЮВАЛЬНОЇ КОНДЕНСАЦІЇ МЕТАНУ	ДВ Шаган, ЛМ Бугаєва	МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ Черкаський національний університет ..., 0		0	0.7524	[відсутність співавтора в КПІ][одна публікація в джерелі або джерело не розпізнане]
Системне моделювання динаміки регіональних процесів у контексті сталого розвитку	ЛМ Бугаєва, ММ Османов, ГО Статюха	НТУУ «КПІ», 2010	2010	0	0.99	

Рисунок 2.28 – Приклад профілю з великою кількістю успішно пройдених перевірок

Також наведемо приклад публікації з профілю Савицької О.М., якій було помилково додано публікацію Савицької О.С. даний приклад представлено на рис. 2.29.

Розвиток обдарованості учнів на уроках трудового навчання в умовах допрофільної підготовки	ОС Савицька	Науковий часопис НПУ імені МП Драгоманова. Серія 5: Педагогічні науки ..., 2015	2015	0	0.405	[Неспівпадіння ініціалів][одна публікація в джерелі або джерело не розпізнане]
--	-------------	---	------	---	-------	--

Рисунок 2.29 – Приклад помилково доданої публікації і її низький коефіцієнт довіри

Таким чином, можна стверджувати що перевірки почали більш точно відображати рівень довіри до публікації. Кінцеві критерії та їх коефіцієнти впливу наведені у таблиці 2.2.

Таблиця 2.2 – Початкові та кінцеві критерії та їх показники впливу

№	Критерій	Початкові значення критеріїв	Кінцеві значення критеріїв
1.	Не співпадіння прізвища	0.25x	0.25x
2.	Не співпадіння ініціалів	0.6x	0.75x
3.	Відсутність співавтора у ЗВО	0.8x	0.95x
4.	Відсутність публікацій у даному журналі	0.75x	0.8x
5.	Кількість цитувань публікації відрізняється від середньої, без врахування цієї більш ніж на 50%	0.9x	0.9x
6	Індекс Хірша без врахування даної публікації відрізняється від індексу Хірша більш ніж на 50%	0.8x	Вилучено
7.	Кількість зв'язків у вершини графу співавторства науковця, де n – кількість пов'язаних нод. При цьому максимальне значення n не перевищує 10	$0.05 * 2nx$	$((0.01 * n) + 0.9)x$

Розглянемо граничні значення рішень, які приймаються згідно показника коефіцієнту довіри.

Так, як стверджувалось раніше, помилково доданими вважались публікації з показником коефіцієнту довіри менше 0.3, потенційно

ненадійними публікації з коефіцієнтом довіри в межах 0.3 – 0.8, та коректними публікації з коефіцієнтом довіри більше за 0.8. З огляду на результати окремих перевірок пропонується змінити ці значення та встановити їх на наступному рівні.

Помилково додані публікації – 0.3. Даний показник залишається таким самим, такий низький рівень коефіцієнту довіри показує, що багато перевірок не були пройдені та таким публікаціям має бути приділена особлива увага.

Потенційно ненадійні – [0.3, 0.7] – в цих межах є коефіцієнт довірити потенційно ненадійних публікацій, а також таких, які не пройшли більш ніж дві малозначущі перевірки, або більш ніж одну значущу. Такі публікації варто перевірити та підтвердити коректність їх авторства.

Коректні публікації – коефіцієнт довіри більший за 0.7. такі публікації з високою ймовірністю є правильними.

Важливо зауважити, що дані перевірки не є однозначною характеристикою публікацій та несуть суто рекомендаційний характер. Чим нижчий коефіцієнт довіри, тим більша ймовірність, що необхідно звернути увагу на дану публікацію.

Показники критеріїв коефіцієнту надійності на усіх ітераціях наведені у «Додатку А» на «Плакаті 7».

Результати роботи даних перевірок з прикладами та розподілами будуть наведені у наступному підрозділі даної роботи.

2.6 Результати досліджень

За допомогою даних перевірок було оцінено 77 147 публікацій. З них 2 401 отримали показник коефіцієнту надійності 1. Наведемо детальний розподіл коефіцієнту надійності в таблиці 2.3.

Таблиця 2.3 – Кількість публікацій в заданих діапазонах

Діапазон	Кількість публікацій в діапазоні
0.1 - 0.2	12 671
0.2 - 0.3	1 967
0.3 - 0.4	0
0.4 - 0.5	1 116
0.5 - 0.6	9 115
0.6 - 0.7	14 929
0.7 - 0.8	15 466
0.8 - 0.9	7 785
0.9 - 1	8 709
рівне 1	2 407

Зобразимо даний розподіл графічно на рис. 2.30 нижче.



Рисунок 2.30 – Діаграма розподілу коефіцієнту надійності за діапазонами

Представимо розподіл по граничних значеннях, наведених вище, в таблиці 2.4.

Таблиця 2.4 – Розподіл коефіцієнту надійності в діапазонах рішення

Рішення	Діапазон	Кількість публікацій
Потенційно помилкові	<0.3	14816
Потенційно не надійні	0.3-0.6	10231
Коректні	>0.6	51314

Наведемо порівняння отриманих результатів з результатами попереднього дослідження оцінки публікації [7]. Представлений в ній алгоритм оцінив як ненадійні 34 608 робіт, що показувало велику кількість

false-positive спрацьовувань. Оскільки попередня версія алгоритму використовувала двокласову систему прийняття рішень (0-1 результат), для порівняння результатів необхідно об'єднати класи рішень «Потенційно помилкові» та «потенційно ненадійні». Результати порівняння наведено в таблиці 2.5.

Таблиця 2.5 – Порівняння результатів з попереднім дослідженням

Рішення	«Коефіцієнт Надійності»	Попереднє Дослідження
Коректні	51314	42471
Потенційно помилкові	25047	34676

Результати даних порівнянь зображені графічно на рис. 2.31.

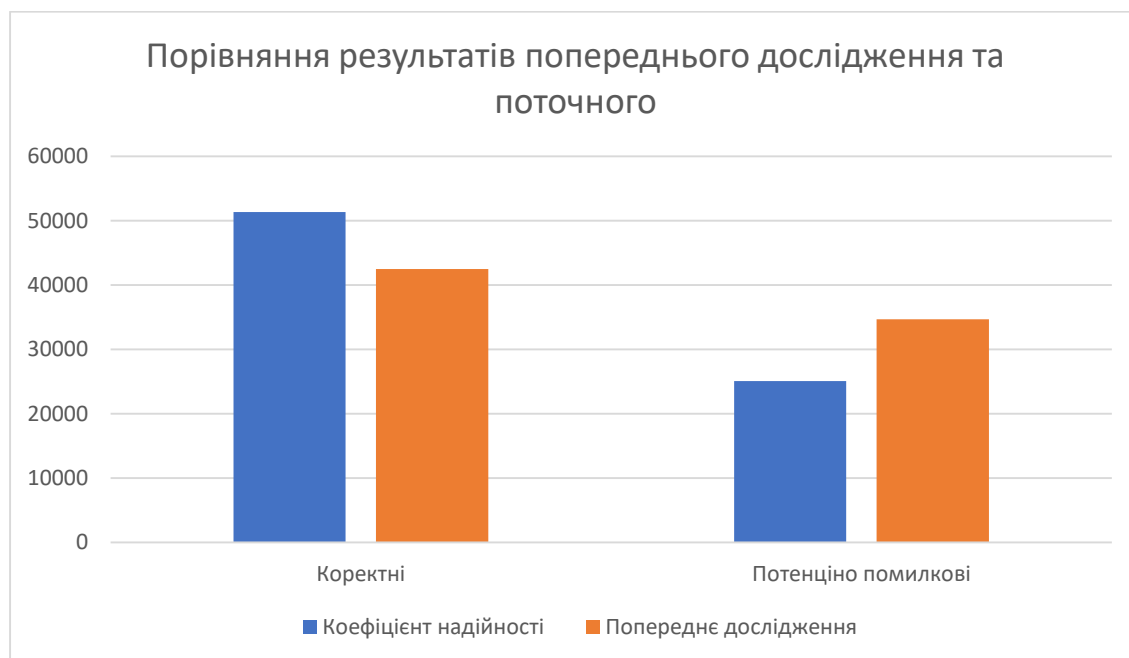


Рисунок 2.31 – Порівняння результатів поточного і попереднього дослідження

Як видно з результатів роботи алгоритму, заснованого на коефіцієнті надійності, оновлений алгоритм показує меншу кількість false-positive спрацьовувань та визначає критерій у певному діапазоні, а не позначає публікацію як ненадійну при визначені помилки. Дослідження, представлене в магістерській дисертації, надає більш гнучку оцінку та показує точніші результати на тому самому наборі вхідних даних.

Висновок до розділу

Даний розділ вводить моделі та методи, які необхідні для визначення помилково доданих публікацій, шляхом розрахунку коефіцієнту надійності. За допомогою описаних моделей і методів виводиться алгоритм розрахунку коефіцієнту надійності публікації.

Впродовж досліджень та випробовувань алгоритму вносилися певні зміни в критерії, на основі яких здійснюється оцінка коефіцієнту надійності, та в коефіцієнти впливу цих критеріїв на коефіцієнт надійності. Фінальним алгоритмом було оцінено більше 70 тисяч публікацій наукових співробітників ЗВО.

В розділі 2.1 наведено змістовна та узагальнена постановку задачі та цілі, що мають бути досягнуті.

Розділ 2.2 представляє математичні моделі, на яких формуються алгоритми, необхідні для досягнення цих цілей.

Розділ 2.3 містить огляд наявних досліджень на дану тему та огляд літератури, використаної в межах цієї роботи.

Розділ 2.4 наводить методи та алгоритми, які використовуються для розрахунку коефіцієнту довіри до публікації.

У розділ 2.5 описано процес проведення розробки та оптимізації алгоритму розрахунку коефіцієнту довіри, а також аналіз алгоритму та

процес визначення його окремих критеріїв та їх показників впливу для досягнення поставлених цілей.

В розділі 2.6 наведені узагальнені результати досліджень та порівняння з попередніми дослідженнями. Як показують результати, було досягнуто значного покращення точності визначення та зменшення кількості false-positive результатів.

У результаті даної роботи усі поставлені цілі у розділі 2 були виконані. Результати описані в пункті 2.6.

3 РОЗРОБКА ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

Даний розділ описує технічні особливості розробленої інформаційної підсистеми. В даному розділі наводяться засоби розробки та технології, що були використані для реалізації веб застосунку, що виконує поставлену мету даної дисертації. Розділ наводить архітектурні рішення за допомогою формального опису та UML діаграм. Наводиться інструкція користувача даної підсистеми та діаграма розгортання.

3.1 Засоби розробки

Підсистема представляє собою легковісний, сучасний веб застосунок. Основними мовами реалізації даного веб-застосунку є PHP 8.2 для backend частини та JavaScript для frontend частини. Backend частина веб-застосунку заснована на сучасному фреймворку Symfony.

Нижче більш детально перераховано основні технології та фреймворки, які були використані для розробки веб-застосунку.

3.1.1 Symfony

В основі backend частини застосунку є сучасний, легкий для підтримки PHP фреймворк Symfony. Symfony фреймворк дозволяє створювати як повноцінні веб застосування, так і API застосунки. Дане веб застосування є певною сумішшю – фреймворк відповідає за переходи між сторінками та передає дані у відповідь на API запити від frontend частини застосунку.

Фреймворк дозволяє використовувати вбудовані компоненти для готових реалізацій окремих функцій системи, таких як авторизація, функції безпеки, маршрутизації, відправки листів та сповіщень тощо.

Symfony дозволяє створювати застосунки у архітектурі MVC.

Архітектура MVC дозволяє розмежувати основні функції на більш незалежні архітектурні складові. Розглянемо кожен з них та їх реалізації в рамках фреймворку більш детально.

View в даному фреймворку реалізовується за допомогою гнучкої системи шаблонування. Шаблони фреймворку Symfony мають назву Twig та представляють собою параметризовані html файли, в які можна передавати дані з контролерів для подальшого відображення та взаємодії з користувачем.

Контролери описують точки входу в застосунок. Система маршрутизації фреймворку дозволяє налаштовувати маршрути-запити браузера та задавати їм метод, який буде обробляти ці запити та повертати у відповідь шаблон або серіалізовану відповідь у разі використання API запиту.

Модель представляє собою сутності, які пов'язані з базою даних. По суті, об'єкт сутності є представленням запису в базі даних. Система форм дозволяє створювати, заповнювати та змінювати сутності даними з бази або форми сторінки з подальшим збереженням в базі даних.

Додатковою перевагою даного фреймворку є можливість виконувати Команди. Команда – це частина програми, яка виконується не безпосередньо через дії користувача, а запускається адміністратором через термінал. Перевагою Команд є можливість непомітно для користувача вносити зміни в існуючі дані, запускати великі процеси на фоні, або налаштовувати cron-команди, які будуть запускати автоматично з заданим інтервалом.

3.1.2 Doctrine

Важливою складовою фреймворку Symfony є можливість налаштувати підключення до бази даних та працювати з записами бази даних, як з об'єктами. Це дозволяє робити ORM Doctrine. ORM – Object-Relational Mapping, або об'єктно-реляційне співставлення, є архітектурним підходом у програмуванні, який пов'язує базу даних і об'єктне представлення сутностей, що відображаються таблицями цієї бази даних.

Так Doctrine за допомогою простих команд дозволяє взаємодіяти з базою даних, підтримувати її, вносити зміни в структуру таблиць бази даних, змінюючи лише відповідні сутності.

Також Doctrine дозволяє встановлювати правила валідації даних, наприклад за довжиною або умовою.

Загалом, Doctrine дає змогу легко поєднати веб застосунок на PHP та базу даних.

3.1.3 VueJS

VueJS є сучасним JavaScript фреймворком для створення веб-інтерфейсів. Основними перевагами фреймворку є простота, гнучкість та легкість, що робить його гарним варіантом для швидкого створення інтерактивних веб-застосунків різної складності.

Однією з ключових особливостей є використання реактивного пов'язування даних. Це означає автоматичне відображення внесених змін, без необхідності оновлення сторінки.

На додаток до основної бібліотеки, VueJS надає екосистему плагінів та розширень, таких як додаткові бібліотеки для управління станом застосунку, маршрутизації, або додаткові елементи відображення, таких як нові випадаючі списки та інше.

3.1.4 API

В даному застосунку VueJs використовується для динамічного відображення сторінок та завантаження даних на сторінки, а також для реалізації пошуку. Для отримання даних в описаних сценаріях, VueJS має здобути їх і лише їх від сервера, без рендеру сторінки. Для цього застосовуються API (Application Programing Interface). Це набір методів та інструкцій, що використовуються для взаємодії різних програмних модулів однієї системи, або між різними системами загалом для обміну даними. Загальна схема API архітектури зображена на рис. 3.1.

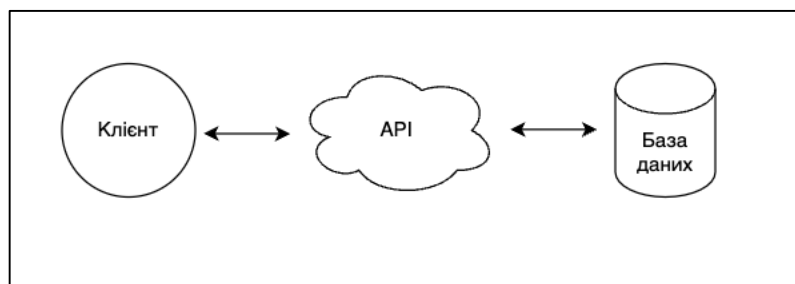


Рисунок 3.1 – Загальна схема API архітектури

Загалом, API – це програмний інтерфейс, за допомогою якого частина програми на frontend стороні отримує дані з бази від backend сторони. В контексті даного застосунку – для отримання, наприклад, інформації про список факультетів або результати пошуку. Програмні Інтерфейси мають різні призначення, основні з яких:

- інтеграція – Різні частини однієї інформаційної системи або різні інформаційні системи можуть взаємодіяти між собою;
- автоматизація – API дозволяють автоматизувати процеси отримання та відправки даних;

- масштабованість – Оскільки API дозволяють легко інтегруватись в інші сервіси та підсистеми, вони дозволяють легше масштабувати інформаційні системи.

3.1.5 MySQL

Для збереження даних використовується сучасна широко використовувана база даних MySQL. MySQL є реляційною системою керування базою даних (СКБД), що часто використовуються для побудови веб додатків та зберігання даних в них. Особливостями даної СКБД є:

- масштабованість – MySQL дозволяє легко розширювати існуючі таблиці, створювати нові та пов'язувати їх з існуючими;
- безпека – MySQL має вбудовані безпекові інструменти, такі як шифрування, розмежування доступу та інші;
- відкритий вихідний код – Дана база даних не потребує оплати для її використання і є гарним рішенням для невеликих застосунків.

Загалом, MySQL є потужною СКБД, яка дозволяє швидко та легко розгортати бази даних для застосунків різною складності та різних розмірів.

3.1.6 Bootstrap

Для реалізації користувацького інтерфейсу було використано сучасний frontend фреймворк Bootstrap. Bootstrap є легким для підтримки frontend фреймворком для розробки та дизайну швидких та зручних для використання веб сторінок. Bootstrap це набір HTML, CSS та JavaScript бібліотек, які можуть бути швидко інтегровані в існуючі додатки, або використані як інструмент для створення нових з нуля. Bootstrap має гарну проєктну документацію та широкий набір сторонніх користувацьких бібліотек для використання. Bootstrap має велику кількість навчальних матеріалів, прикладів використання компонентів та набори компонентів

(теми), необхідних для реалізації вимог інформаційних систем. Також Bootstrap дозволяє стандартизувати код застосунку, що полегшує інтеграцію з іншими бібліотеками та компонентами.

Основними перевагами Bootstrap є:

- зручний дизайн – Bootstrap дозволяє створювати зручні та адаптивні сторінки, які будуть виглядати відповідно до розмірів екрану і підлаштовуватись під мобільні пристрої;
- підтримка Браузерів – Bootstrap підтримує усі сучасні браузерери та дозволяє розроблювати сторінки, які матимуть однаковий вигляд в усіх браузерах.

3.1.7 Composer

Для завантаження та встановлення додаткових бібліотек та пакетів для мови PHP використовується проєктний менеджер Composer. Даний інструмент спрощує процес управління залежностями та бібліотеками й гарантує, що необхідні пакети будуть автоматично встановлені при розгортанні застосунку.

В основі Composer є файл `composer.json`, в якому визначаються необхідні бібліотеки та їх версії, а також відомості про проєкт в цілому. Composer підтримує можливість обмеження версій для пакетів, таким чином можна задати конкретні версії, які використовуватимуться в проєкті.

Також Composer може автоматично слідкувати за необхідністю оновлення версій та дозволяє швидко оновити версію конкретного пакету, або всіх пакетів проєкту в цілому до останньої версії, або до відповідного обмеження версії, заданої в проєкті.

Composer спрощує управління залежностями у PHP-проєктах та забезпечує встановлення та оновлення необхідних залежностей. Він є

стандартним інструментом розробки PHP, що широко використовується PHP-розробниками.

Основними перевагами Composer є:

- стандартизація – Composer дозволяє стандартизувати процес встановлення та обміну бібліотеками та пакетами для усіх проєктів на PHP;
- автоматизація – Composer автоматично завантажує необхідні для проєкту дані та бібліотеки, які описані у файлі `composer.json`;
- контроль версій – Composer дозволяє слідкувати та оновлювати пакети до нових версій за необхідності.

3.1.8 Webpack

Webpack є широко використовуваним інструментом для веб розробки з використанням JavaScript. Він дозволяє збирати складні проєкти з багатьох модулів та бібліотек в один JS файл, який буде виконуватись на frontend частині застосунку.

Webpack використовує графи залежностей шляхом аналізу залежностей між модулями в проєкті. Даний граф дозволяє об'єднати вихідний код в єдиний файл, що буде містити весь необхідний код та ресурси для виконання.

При цьому Webpack здатний також обробляти CSS файли та інші ресурси проєкту.

Webpack дозволяє легко розширювати та масштабувати проєкт за допомогою багатьох можливостей, таких як:

- гаряча заміна модулів, що дозволяє підміняти модулі без необхідності перезавантаження сторінки;
- підключення плагінів для оптимізації вихідного коду, таких як його розбиття, мінімізація та інше;

- розділення коду, що дозволяє зменшити час завантаження шляхом завантаження лише частини коду;
- тонкі налаштування за допомогою окремих плагінів та завантажувачів, що дозволить налаштувати збір файлів згідно проєктних вимог.

Загалом Webpack є потужним інструментом, що може покращити розробку та пришвидшити і спростити процес розгортання інформаційних систем.

3.2 Архітектура програмного забезпечення

Даний веб застосунок реалізований за допомогою архітектурного шаблону MVC, тому для наведення архітектури програмного забезпечення необхідно буде показати структуру класів на різних рівнях архітектури.

Модель в даній архітектурі представляє собою шар класів-сутностей та їх класів-репозиторіїв для взаємодії з базою даних.

Наведемо діаграму, що описує дані класи та їх зв'язки на рис. 3.2.

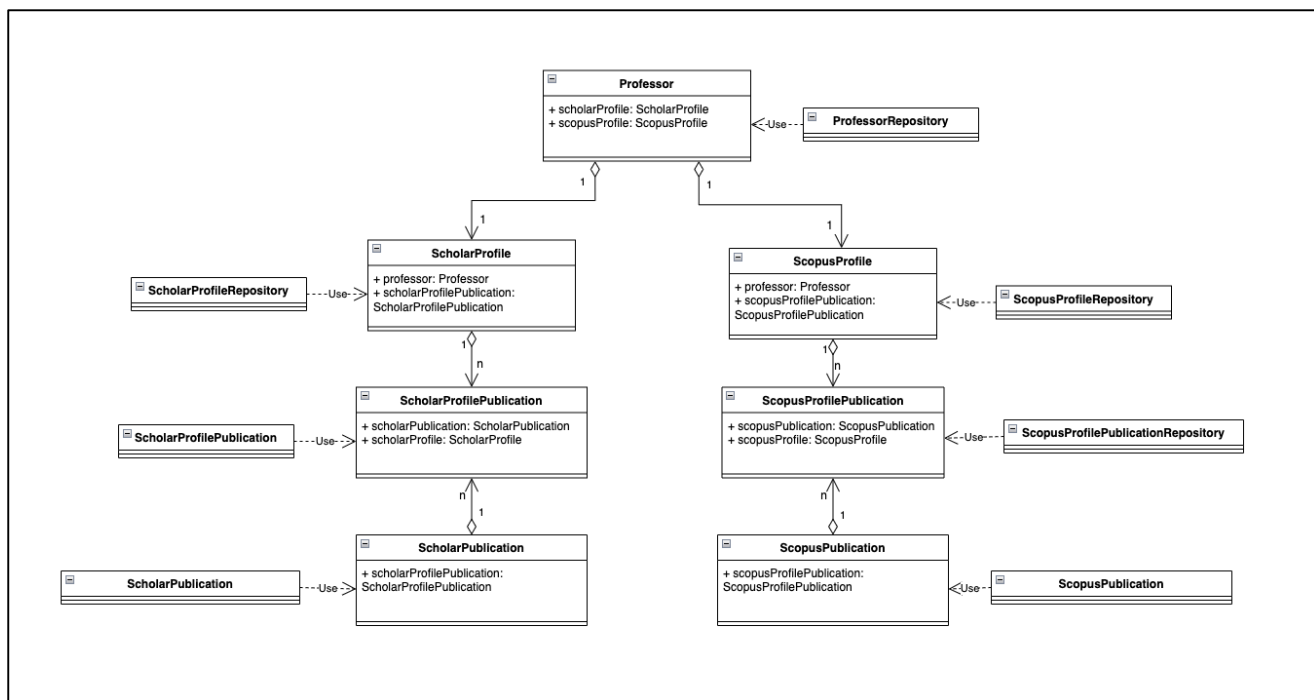


Рисунок 3.2 – Діаграма класів сутностей та їх репозиторіїв

На даній діаграмі вказані лише класи та їх поля що використовуються для зв'язку з іншими класами. Повний набір полів даних класів-сутностей можна переглянути у «Додатку А» на «Плакати 3» даної роботи.

Об'єктно-реляційне співставлення копіює структуру таблиць бази даних у відповідні об'єкти та робить клас-репозиторій, що реалізовує інтерфейс для доступу до цієї таблиці. В загальному, класи-сутності та їх зв'язки представлені на рисунку 3.2.

Дана діаграма представляє загальний опис зв'язків між сутностями.

Для взаємодії користувача та серверу використовуються контролери. Контролери використовують сервіси для обробки та отримання даних. Діаграма класів, що показує контролери та сервіси наведена в «Додатку А» на «Плакати 5».

Наведемо приклад виконання та обробки запиту контролером на діаграмах послідовностей. Наведемо приклад відкриття сторінки перегляду кафедр факультетів. Діаграма послідовності представлена у «Додатку А» на

«Плакати 6». На діаграмі зображено користувача-ініціатора апі запиту на отримання кафедр факультетів.

Визначений метод контролера приймає запит за задане посилання та приймає як параметр методу строкове представлення назви факультету.

Контролер використовує DataService для отримання та обробки даних. Контролер викликає відповідний метод та передає назву факультету. Сервіс робить запит на отримання репозиторію відповідної сутності та формує DQL запит на отримання кафедр певного факультету.

Після отримання даних DataService обробляє отримані дані, перераховує індекс Хірша та повертає відсортовані за індексом Хірша об'єкти контролеру. Так як контролер може повертати json представлення відповіді на запит, використовується SerializeService для серіалізації (для перетворення ООП-представлення сутності на масив) отриманих даних.

Отриманий серіалізований масив повертається користувачу для обробки на frontend частині застосунку і подальшого відображення.

Так як інші АПІ запити мають аналогічну структуру, обмежимося наведенням лише одного апі запиту у вигляді діаграми послідовності.

Представимо у загальному вигляді діаграму компонентів на рис. 3.3.

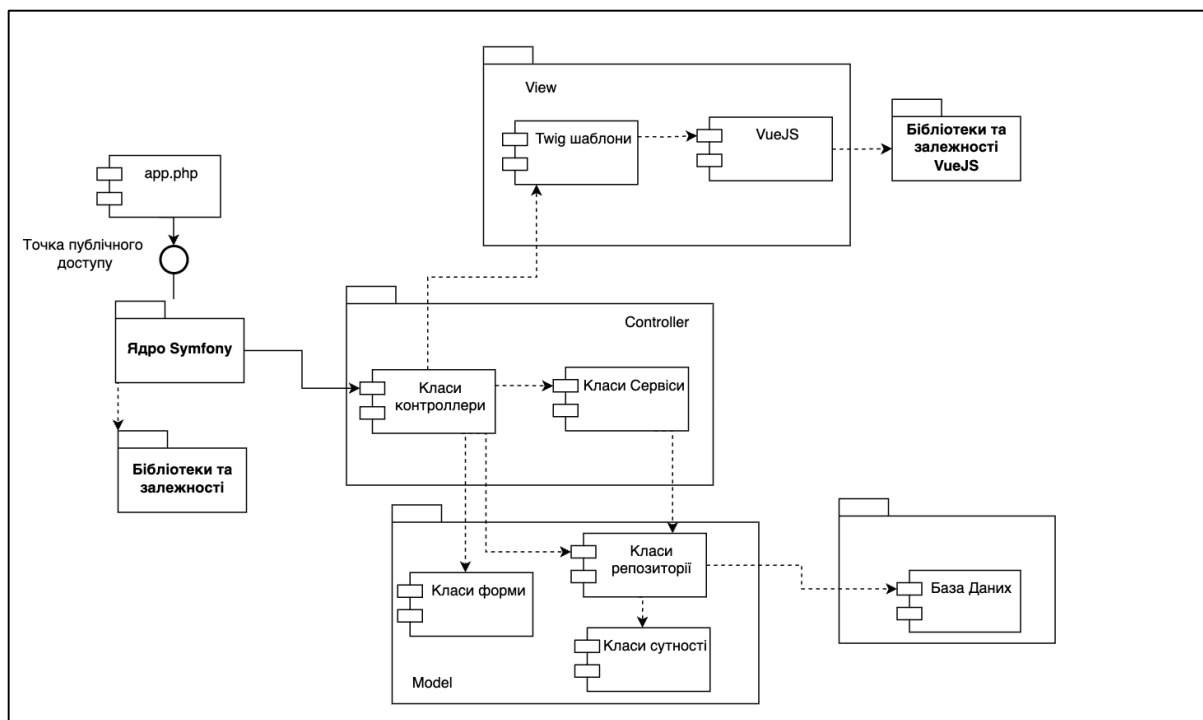


Рисунок 3.3 – Діаграма компонентів

Як видно з діаграми, користувач взаємодіє з контролером через веб сервіс. Контролери мають доступ використовують сервіси для пошуку та обробки даних. Сервіси використовують репозиторії для доступу до даних, які зберігаються у базі даних.

Дана структура компонентів є рекомендованою до використання у даному фреймворку та дозволяє більш явно розділяти застосунок на шари для реалізації архітектурного шаблону MVC.

Наведені вище діаграми в описують структуру та архітектуру представленої інформаційної системи, описують структуру класів, показують зв'язки між сутностями та демонструють типову логіку обробки запиту на діаграмі послідовностей.

3.3 Інструкція користувача

Підсистема представлена веб-застосунком. Для роботи з нею немає необхідності встановлювати на персональний комп'ютер або інший пристрій додаткове програмне забезпечення.

Для роботи з представленою інформаційною системою необхідно мати будь який сучасний браузер останньої версії та стабільне підключення до мережі інтернет.

На даних прикладах буде використовуватись локальний сервер застосунку.

При відкритті головної сторінки, перед користувачем буде індекс-сторінка застосунку – список факультетів та загальна інформація про них. Дана сторінка представляє варіант використання «Перегляд інформації про факультети» з рис. 1.1. На даній сторінці представлена таблиця, колонками якої є – назва факультету, загальна кількість статей у наукових співробітників факультету, загальна кількість цитувань цих статей та перерахований індекс Хірша для цього факультету.

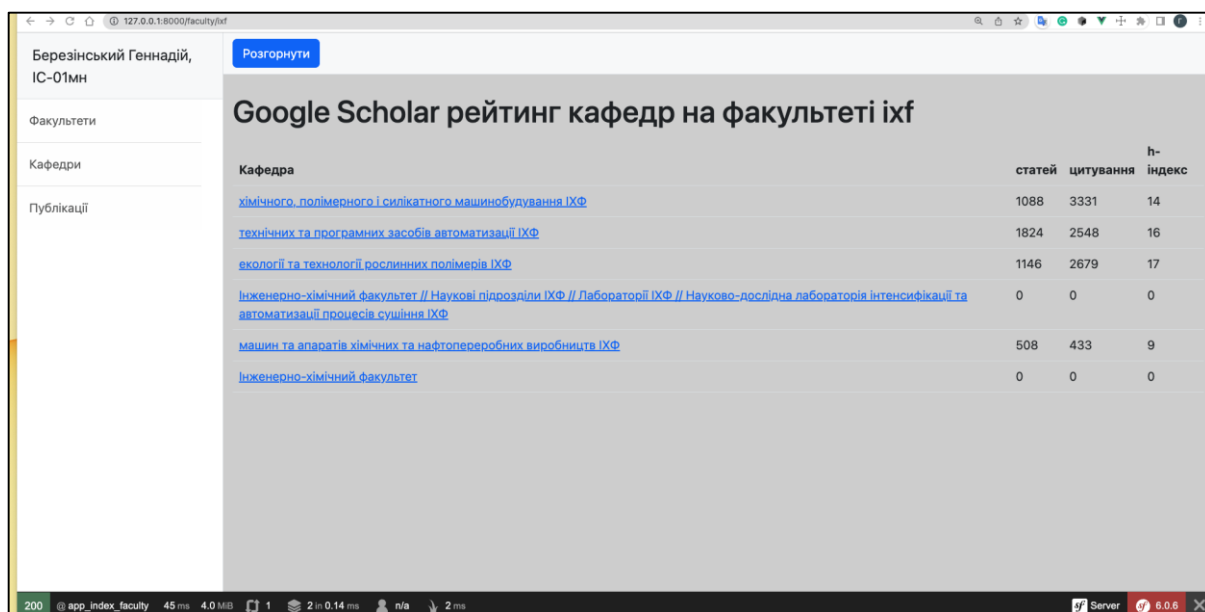
З цієї сторінки можливо перейти на сторінку факультету натиснувши на назву факультету, або за допомогою кнопок на бічній панелі на список кафедр або на вікно пошуку публікацій. Дана сторінка представлена на рис. 3.4.

факультет	статей	цитування	h-індекс
fse	3488	7822	34
rtf	1703	3420	20
its	2376	4818	18
ixf	4566	8991	23
fom	1478	4004	20
xtf	4139	16484	37
fmm	5861	14578	37
zf	837	1463	15
iasa	1909	14259	31
fmf	4944	20682	44
fi	4126	3414	20
iae	4125	7057	22

Рисунок 3.4 – Головна сторінка інформаційної системи

Обравши сторінку факультету можна перейти на сторінку факультету і переглянути кафедри факультету з загальною інформацією про них. На даній сторінці представлено таблицю, колонками якої є назва кафедри, сумарна кількість статей наукових співробітників, з цієї кафедри, їх загальна кількість цитувань та індекс Хірша для даної кафедри. Дана сторінка представляє варіант використання «Перегляд інформації про кафедри», представленої на рис. 1.1.

З даної сторінки можливо перейти на сторінку кафедри, натиснувши на назву кафедри або на сторінку факультетів, перегляд списку всіх кафедр чи пошуку публікації через бічну панель. Дана сторінка наведена на рис. 3.5.



Кафедра	статей	цитування	h-індекс
хімічного, полімерного і силікатного машинобудування IXФ	1088	3331	14
технічних та програмних засобів автоматизації IXФ	1824	2548	16
екології та технології рослинних полімерів IXФ	1146	2679	17
Інженерно-хімічний факультет // Наукові підрозділи IXФ // Лабораторії IXФ // Науково-дослідна лабораторія інтенсифікації та автоматизації процесів сушіння IXФ	0	0	0
машин та апаратів хімічних та нафтопереробних виробництв IXФ	508	433	9
Інженерно-хімічний факультет	0	0	0

Рисунок 3.5 – Сторінка перегляду факультету

Обравши сторінку кафедри можна переглянути список наукових співробітників, які пов'язані з даною кафедрою. Дана сторінка представлена у вигляді таблиці, колонками якої є Ім'я наукового співробітника, кількість його публікацій, кількість цитувань та індекс Хірша наукового співробітника. Дана сторінка представлена на рис. 3.6.

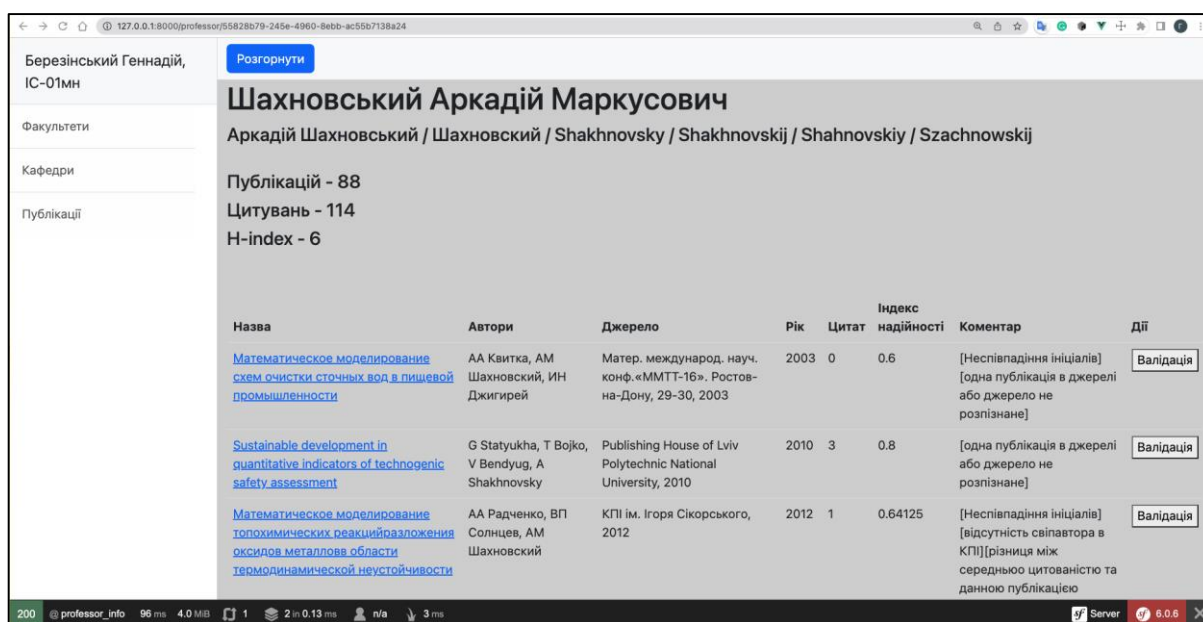


Ім'я	статей	цитування	h-індекс
Бойко Тетяна Владиславівна	100	316	9
Джигирей Ірина Миколаївна	100	226	8
Безносик Юрій Олександрович	100	209	7
Бугаєва Людмила Миколаївна	76	95	7
Шахновський Аркадій Маркусович	88	114	6
Бендюг Владислав Іванович	72	110	6
Абрамова Алла Олександрівна	69	113	6
Квітка Олександр Олександрович	67	146	6
Жученко Анатолій Іванович	65	122	6
Лукінок Михайло Васильович	20	78	6
Жученко Олексій Анатолійович	100	115	6
Кваско Михайло Зіновійович	30	83	5
Коржик Михайло Володимирович	52	62	5
Савицька Ольга Вікторівна	06	01	5

Рисунок 3.6 – Сторінка кафедри

З даної сторінки можна перейти на сторінку наукового співробітника обравши його зі списку або на список факультетів, список кафедр чи пошук публікацій через бічне меню.

Якщо було обрано сторінку наукового співробітника, можна переглянути інформацію про наукового співробітника. На даній сторінці можна переглянути загальну інформацію про наукового співробітника, ім'я вказане в профілі Google Scholar, інформацію про кількість його публікацій, індекс Хірша та список публікацій з інформацією про них. Дана сторінка представлена на рис. 3.7.



Назва	Автори	Джерело	Рік	Цитат	Індекс надійності	Коментар	Дії
Математическое моделирование схем очистки сточных вод в пищевой промышленности	АА Квитка, АМ Шахновский, ИН Джигирей	Матер. междунар. науч. конф. «ММТТ-16». Ростов-на-Дону, 29-30, 2003	2003	0	0.6	[Неспівпадіння ініціалів] [одна публікація в джерелі або джерело не розпізнане]	Валідація
Sustainable development in quantitative indicators of technogenic safety assessment	G Statyukha, T Bojko, V Bendyug, A Shakhnovsky	Publishing House of Lviv Polytechnic National University, 2010	2010	3	0.8	[одна публікація в джерелі або джерело не розпізнане]	Валідація
Математическое моделирование топочимических реакций разложения оксидов металлов в области термодинамической неустойчивости	АА Радченко, ВП Солнцев, АМ Шахновский	КПІ ім. Ігоря Сікорського, 2012	2012	1	0.64125	[Неспівпадіння ініціалів] [відсутність свіпавтора в КПІ][різниця між середньою цитованістю та данною публікацією]	Валідація

Рисунок 3.7 – Сторінка наукового співробітника

З сторінки наукового співробітника є можливість переглядати публікації та інформацію про них, запустити валідацію даної публікації за допомогою кнопки Валідація. Сторінка є реалізацією варіанту використання «Перегляд інформації про наукових співробітників».

З даної сторінки є можливість перейти на сторінку публікації натиснувши на неї, або на список факультетів, список кафедр чи пошук публікацій на бічній панелі.

Розглянемо сторінку публікації. Дана сторінка зображена на рис. 3.8.

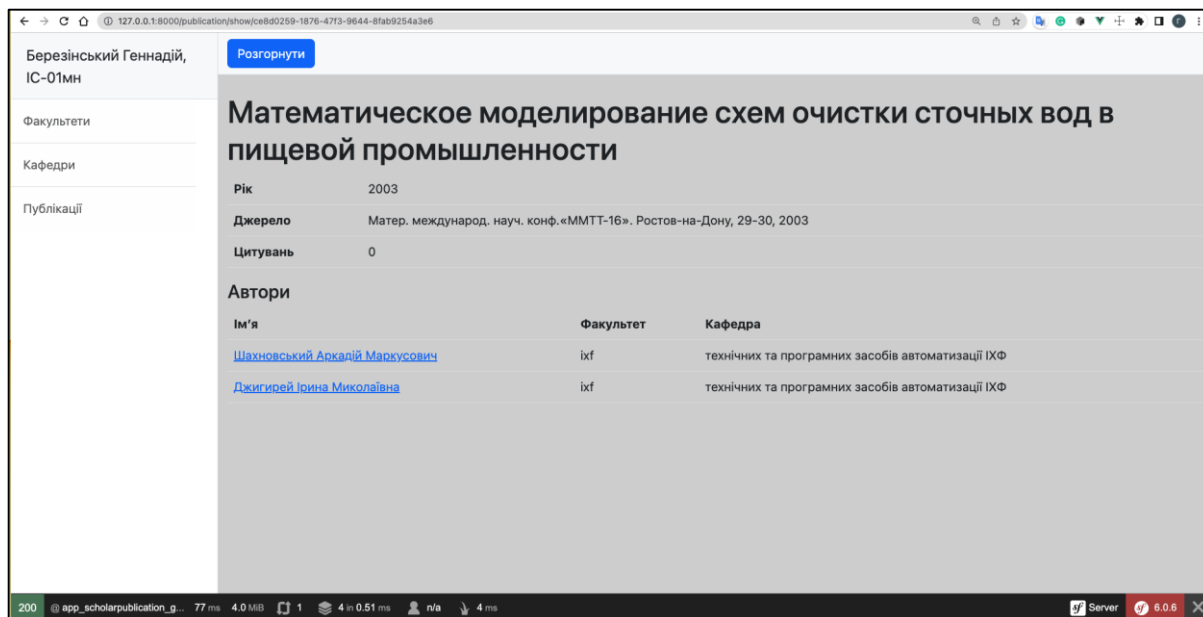


Рисунок 3.8 – Сторінка публікації

Дана сторінка показує загальну інформацію про публікацію та її авторів.

З даної сторінки можливо перейти на сторінку наукового співробітника або список факультетів, список кафедр чи пошук публікацій.

Сторінка пошуку публікацій дозволяє шукати публікації за назвою, авторами, видавництвом або роком публікації. З даної сторінки можна перейти на сторінку публікації або через бічне меню на список факультетів, кафедр.

Зобразимо на рис. 3.9-3.12 можливості пошуку по різному параметру.

Застосунок самостійно визначає запит пошуку та робить пошук, додаткових параметрів обирати не потрібно.

Березінський Геннадій, ІС-01мн

Факультети

Кафедри

Публікації

Розгорнути

Google Scholar пошук публікацій

2014

Назва	Автори	Джерело	Рік	Цитат
Моделювання та прогнозування соціально-економічного розвитку регіонів України	АІ Сухоруков	Стратегічні пріоритети, 199–201, 2014	2014	122
Diamondoids: functionalization and subsequent applications of perfectly defined molecular cage hydrocarbons	MA Gunawan, JC Hierso, D Poinsot, AA Fokin, NA Fokina, BA Tkachenko, ...	New Journal of Chemistry 38 (1), 28–41, 2014	2014	93
Основи охорони праці	КН Ткачук, ВВ Зацарний, ДВ Зеркалов, ОІ Полукаров, ВС Козяков, ...	Основа, 2014	2014	84
From the time series to the complex networks: The parametric natural visibility graph	IV Bezudnov, AA Snarskii	Physica A: Statistical Mechanics and its Applications 414, 53–60, 2014	2014	62
Fatous lemma for weakly converging probabilities	EA Feinberg, PO Kasyanov, NV Zadoianchuk	Theory of Probability & Its Applications 58 (4), 683–689, 2014	2014	53
Аналіз сталого розвитку: глобальний і регіональний контексти, Україна в індикаторах сталого розвитку (2013)	А Болдак, С Войтко, О Гавриш, І Джигирей, М Згуровський	НТУУ «КПІ» 2, 172, 2014		49
Limit theorems for multi-indexed sums of random variables	O Klesov	Springer, 2014	2014	36

200 @ publication_list 109 ms 4.0 MB 9 2 in 0.16 ms n/a 4 ms Server 6.0.6

Рисунок 3.9 – Пошук по року, запит пошуку – 2014

Березінський Геннадій, ІС-01мн

Факультети

Кафедри

Публікації

Розгорнути

Google Scholar пошук публікацій

молодий вче

Назва	Автори	Джерело	Рік	Цитат
Навчання та розвиток персоналу в контексті конкурентоспроможності підприємства	ВВ Красношапка, АО Коваленко	Молодий вчений, 162–165, 2014	2014	10
Досвід впровадження електронного урядування в Естонії та його імплементація в Україні	ЕО Архипова, НО Дмитренко	Молодий вчений, 148–152, 2015	2015	7
Функції пам'яті та уваги у студентів навчального відділення плавання НТУУ КПІ	ОГ Черевичко	Молодий вчений, 622–625, 2016	2016	7
Ways for improvement selectivity of semiconductor gas sensors	SV Nagirnyak, TA Dontsova	Молодий вчений, 15–17, 2015	2015	5
Розробка стратегії розвитку підприємства	МВ Шашина, ТА Степаненко	Молодий вчений, 131–135, 2014	2014	4
Стратегія енергетичної безпеки України: виклики та можливості	УВ Бігун, ОО Охріменко	Молодий вчений, 89–92, 2015	2015	4
Англомовний публіцистичний дискурс як соціокультурологічний лінгвістичний феномен сучасного комунікативного простору в перекладацькому аспекті	СВ Федоренко, ТІ Станішевська	Молодий вчений, 194–196, 2015	2015	4
Імітаційне моделювання виробничого процесу складання	АО Нестеренко, СП Вислоух	Молодий вчений, 406–409, 2018	2018	4

200 @ publication_list 109 ms 4.0 MB 17 2 in 0.16 ms n/a 4 ms Server 6.0.6

Рисунок 3.10 – Пошук по видавництві, запит пошуку – «Молодий вче» – початок запити «молодий вчений»

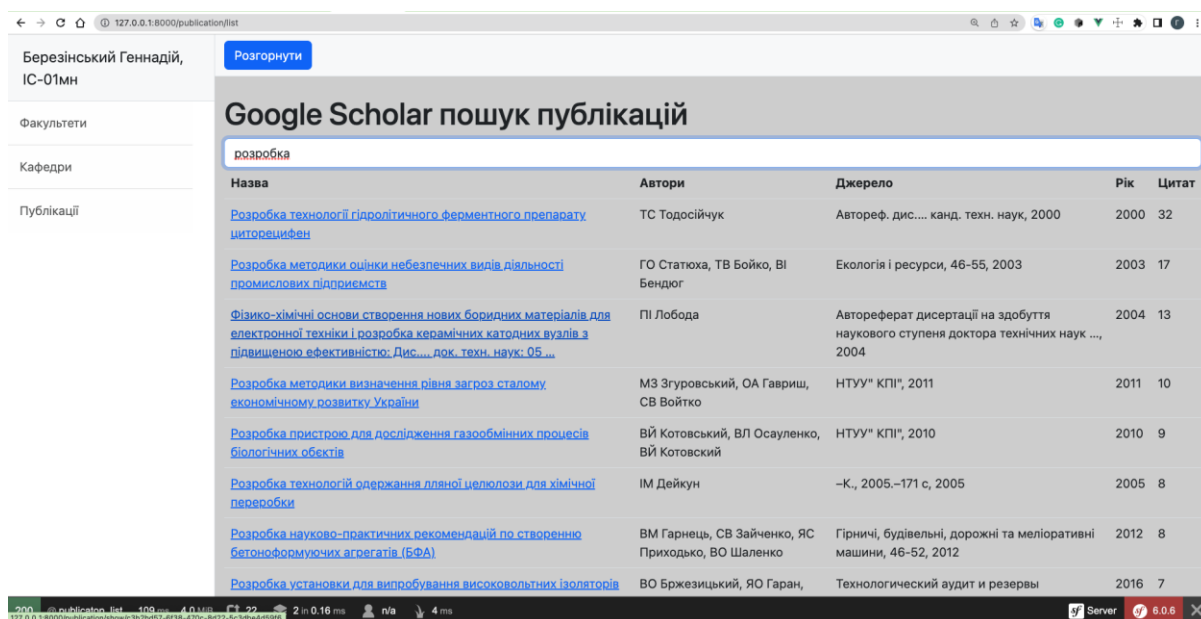


Рисунок 3.11 – Пошук по частині назви, запит пошуку – «Розробка»

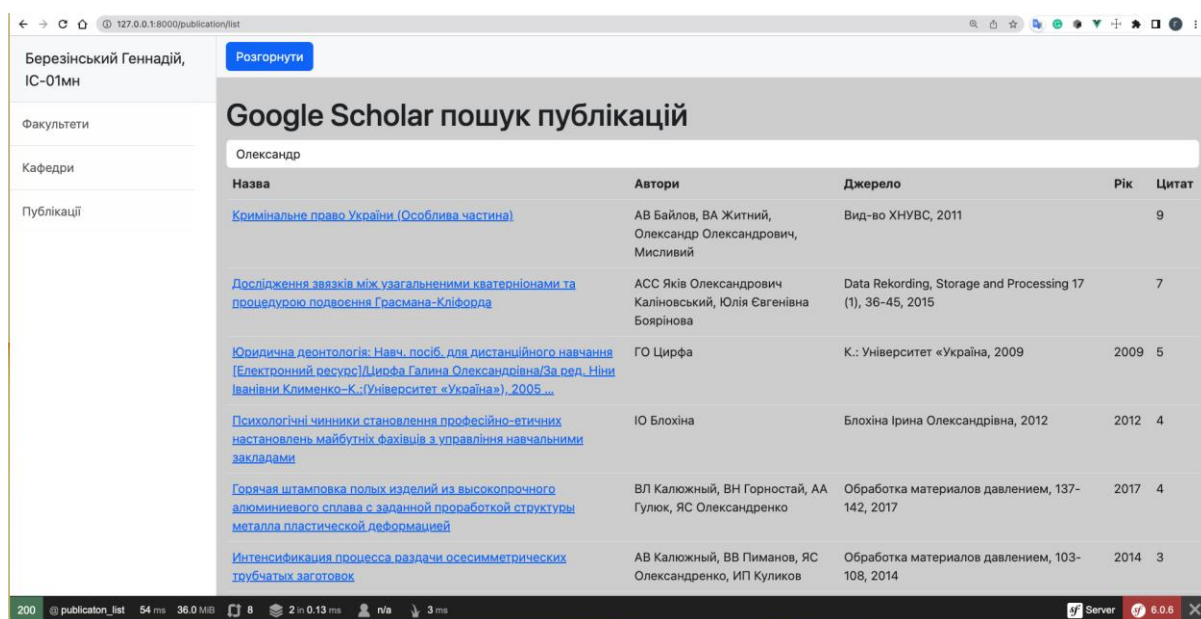


Рисунок 3.12 – Пошук по автору, запит пошуку – «Олександр»

Як видно з рис. 3.12. пошук був здійснений одночасно по назві, авторам, джерелі та році. Сторінка пошуку реалізовує варіант використання «Пошук публікацій».

3.4 Технічне забезпечення

До технічного забезпечення даного застосунку відносяться додаткові сервіси та системи, необхідні для її стабільної роботи.

Для клієнтської сторони застосунку необхідний будь який веб-браузер для мобільних або десктопних пристроїв підтримуваної версії. Доступ до веб-застосунку забезпечується через мережу Інтернет.

Технічне забезпечення для даної системи може бути продемонстровано на діаграмі розгортання, наведеної на рис. 3.13.

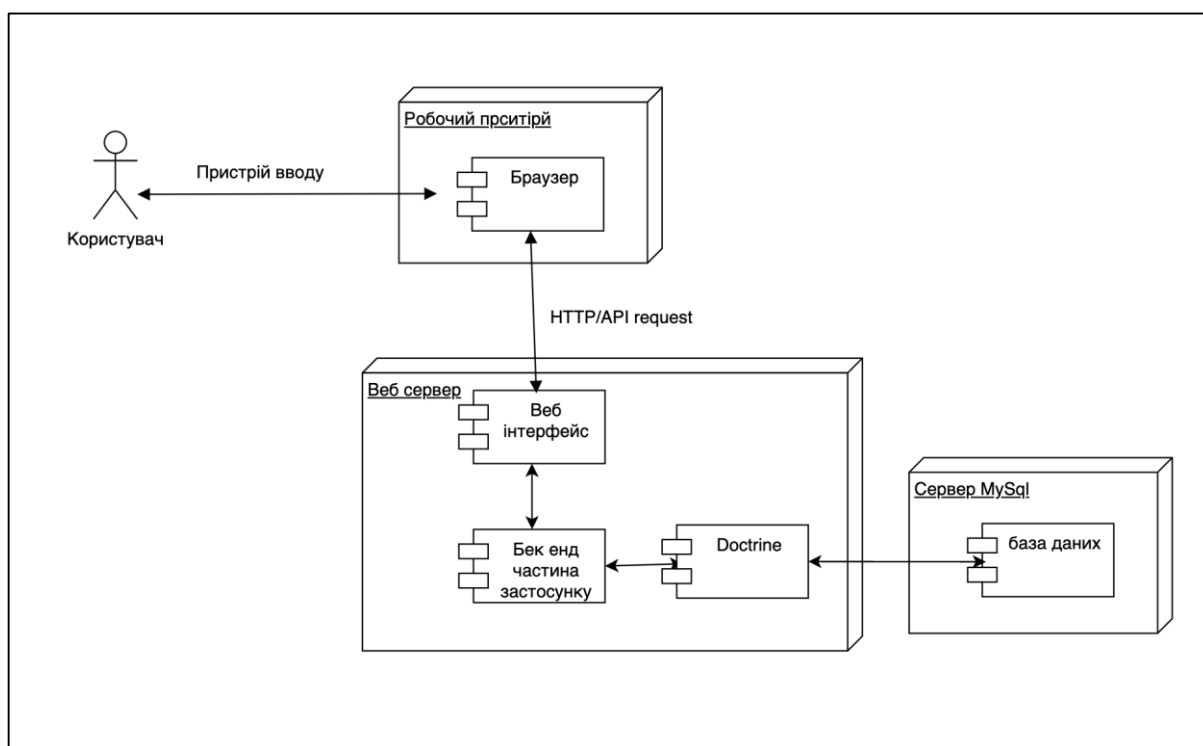


Рисунок 3.13 – Діаграма розгортання

Усі додаткові бібліотеки та пакети встановлюються автоматично при розгортанні на сервері за допомогою команд *composer install* та *npm install*.

Основними компонентами технічного забезпечення інформаційної системи є:

Веб-сервер. Фреймворк дозволяє працювати з різними доступними веб-серверами, втім, рекомендовано використовувати веб-сервер `nginx`. Даний веб-сервер дозволяє швидко розгортати веб-застосунки будь якої складності та не висуває специфічних вимог до апаратного забезпечення.

PHP та JavaScript. Веб-застосування використовує дані мови програмування для своєї роботи. Мають бути встановлені останні актуальні версії даних мов програмування.

База даних. Веб-застосування може використовувати будь-яку базу релятивну базу даних для своєї роботи. Рекомендовано використовувати базу даних `MySQL`, так як саме на ній велась розробка та проводилось тестування. Для стабільної роботи необхідна встановлена остання актуальна версія бази даних, в файлі конфігурацій веб застосування необхідно вказати посилання, для встановлення підключення до бази.

Composer. На сервері для розгортання має бути встановлено пакетний менеджер `Composer`, що при виконанні команди `composer install` встановить усі необхідні додаткові бібліотеки та сервіси для PHP.

NPM. Має бути встановлено пакетний менеджер для JavaScript. При виконанні команди `npm install` буде встановлено усі необхідні пакети та бібліотеки для JavaScript.

ORM. Для зв'язку об'єктів в застосунку та бази даних використовується `ORM Doctrine`. ORM необхідної версії буде встановлена автоматично при запуску команди `composer install` при розгортанні.

Інші програмні пакети та бібліотеки будуть встановлені в процесі розгортання системи автоматично за допомогою описаних пакетних менеджерів.

Висновок до розділу

Даний розділ представляє технічне та технологічне забезпечення магістерської дисертації у вигляді Інформаційної підсистеми.

Наведено список технологій, які використовуються для розробки даної системи та їх короткий опис.

Представлено архітектуру інформаційної системи у вигляді формального опису її та процесів. Також діаграми класів та послідовностей для демонстрації внутрішніх процесів інформаційної системи.

Наводиться докладна інструкція користувача, яка описує можливості інформаційної системи з точки зору кінцевого користувача, та діаграма розгортання інформаційної системи.

Даний розділ описує виконання поставленої мети магістерської дисертації за допомогою розробки інформаційної системи.

ВИСНОВКИ

Дана магістерська дисертація пропонує рішення до вирішення проблеми помилкових публікацій за допомогою оцінки публікацій в профілі на підтвердження авторства.

Представлений метод дозволяє оцінити ймовірність, що публікація дійсно належить автору. Для цього відбувається серія перевірок, кожна з яких модифікує показник коефіцієнту довіри.

Аналіз результатів даного методу показують гарні результати оцінки визначення авторства публікації. Представлений алгоритм перевіряє лише метадані публікації, оскільки не має доступу до повних текстів усіх публікацій.

Магістерська дисертація представлена інформаційною системою, що використовує даний метод для оцінки зібраної інформації про публікацій наукових співробітників ЗВО. Представлено опис технологій та архітектуру розробленої інформаційної системи.

Результатом досліджень в рамках магістерської дисертації є покращення методу визначення помилково доданих публікацій шляхом використання підходу, заснованому на коефіцієнті надійності.

ПЕРЕЛІК ПОСИЛАНЬ

1. Савченко Д. С. Перцептивна модифікація метода N-грам для оцінювання схожості послідовностей символів за коефіцієнтом Серенсена / Д. С. Савченко. // Телекомунікаційні та інформаційні технології. – 2016. – №2. – С. 85–92.

2. Мриглюд О.І. Наукометрія [Електронний ресурс] / О. І. Мриглюд // Енциклопедія Сучасної України / Редкол.: І. М. Дзюба, А. І. Жуковський, М. Г. Железняк [та ін.] ; НАН України, НТШ. – К. : Інститут енциклопедичних досліджень НАН України, 2020. – Режим доступу: <https://esu.com.ua/article-70714>.

3. Постанова Кабінету Міністрів України «Про впорядкування транслітерації українського алфавіту латиницею» від 27 січня 2010 р №55, Київ

4. Google Scholar [Електронний ресурс] – Режим доступу до ресурсу: <https://scholar.google.com.ua/>. — Дата доступу : 12.05.2023.

5. Scopus [Електронний ресурс] – Режим доступу до ресурсу: <https://www.scopus.com/home.uri>. — Дата доступу : 12.05.2023.

6. Закон України "Про вищу освіту" від 01.07.2014 № 1556-VII, стаття 36 "Фінансування закладів вищої освіти"

7. Березінський Г. Аналіз публікацій викладачів закладу вищої освіти на основі відкритих даних / Г. Березінський. // Молодий вчений. – 2021. – №8. – С. 1–6.

8. Мінтій І. С. Огляд наукометричних баз Google Scholar та ORCID / І. С. Мінтій, С. М. Іванова. // Звітна науково-практична конференція Інституту інформаційних технологій. – 2021. – С. 63.

9. Web of Science [Електронний ресурс] – Режим доступу до ресурсу: <https://www.webofscience.com/wos/woscc/basic-search>. — Дата доступу : 12.05.2023.

10. UniRank [Електронний ресурс] – Режим доступу до ресурсу: <https://www.uniranks.com/>. – Дата доступу : 12.05.2023
11. QS [Електронний ресурс] – Режим доступу до ресурсу: <https://www.qs.com/>. – Дата доступу : 12.05.2023.
12. Мамонтов В. Аналіз наукометричних даних науковців на основі графу співавторства / В. Мамонтов, Г. Березінський. // Молодий Вчений. – 2021. – №11. – С. 89–93.
13. Основи дискретної математики Частина III. Основи теорії графів / Г. Г.ШВАЧИЧ, Г. М. БАРТЕНЄВ, О. В. ОНИЩЕНКО, В. В. ТОЛСТОЙ. – Дніпропетровськ: НМетАУ, 2014. – 67 с.
14. BORNMAN L. What do we know about the h index? / L. BORNMAN, H. D. Daniel. // Journal of the American Society for Information Science and technology. – 2007. – №58. – С. 1381–1385.
15. Mitra P. Hirsch-type indices for ranking institutions scientific research output / P. Mitra. // Current Science. – 2006. – №91. – С. 1493.
16. Білощицький, А.О. Аналіз властивостей та функцій метаданих наукометричних суб'єктів у web-просторі / А.О. Білощицький, О.В. Миронов // Управління розвитком складних систем. – 2016. – № 25. – С. 139-143..
17. Одуд О. А. 7 правил для покращення власного наукометричного профілю Google Scholar / О. А. Одуд. // Актуальні питання сучасної інформатики. – 2017. – №5. – С. 390–391.
18. Козицын А. С. Алгоритм разрешения неоднозначности имен авторов в ИАС ИСТИНА / А. С. Козицын, С. А. Афонин. // Современные информационные технологии и ИТ-образованием. – 2020. – №16. – С. 108–117.
19. ВІВЧАРИК В. Особливості методів аналізу текстів для ідентифікації авторства документу / В. ВІВЧАРИК. // Матеріали VII

науково-технічної конференції „Інформаційні моделі, системи та технології. – 2019. – С. 29–29.

20. Srinivasa G. Relevance of innovations in machine learning to scientometrics / G. Srinivasa. // Journal of Scientometric Research. – 2019. – №8. – С. 39–43.

21. Литвин В.В. Метод аналізу авторства тексту на основі стилеметрії / В. В.Литвин, О. В. Оборська, В. А. Висоцька, В. А. Бобик. // Intellectual Systems for Decision Making and Problems of Computational Intelligence . – 2018. – С. 240..

22. Vysotska V. Метод авторифікації тексту науково-технічних публікацій на основі лінгвістичного аналізу коефіцієнтів мовної різноманітності / V. Vysotska. // Radio Electronics, Computer Science, Control. – 2020. – №1. – С. 108–124.

23. Batagelj V. Some analyses of Erdos collaboration graph / V. Batagelj, A. Mrvar. // Social networks. – 2000. – №2. – С. 173–186.

24. Balagura I. V. Дослідження параметрів важливості вузлів у мережах співавторів / I. V. Balagura, D. V. Lande, I. V. Gorbov. // Реєстрація, зберігання і обробка даних. – 2013. – №1. – С. 45–52.

25. Ярмолюк Р. С. Задача аналізу текстових атрибутів в електронному каталозі / Р. С. Ярмолюк. – 2011.

26. КУЗЬМА К. Т. Інформаційна технологія оцінки рівня подібності рядків на основі методу n-грам / К. Т. КУЗЬМА. // ВЧЕНІ ЗАПИСКИ. – 2020. – №31(70). – С. 96–98.

27. Таланчук П. Метод виявлення подібності назв видавництв, стійкий до різних видів скорочень / П. Таланчук, М. Петренко. // Інфокомунікаційні та комп'ютерні технології. – 2021. – №2(02).

28. Minaylo A. Y. Використання відстані Левенштейна для аналізу подібності даних / А. Y. Minaylo, V. A. Turchina. // Питання прикладної математики і математичного моделювання. – 2015.

29. Кузьма К. Обчислювальна технологія нечіткого пошуку в тексті або словнику / Катерина Кузьма. // Геометричне моделювання та інформаційні технології. – 2018. – №2. – С. 65–68.

30. Левенштейн В. И. Двоичные коды с исправлением выпадений, вставок и замещений символов / В. И. Левенштейн. // Доклады Академии наук. – 1965. – №163. – С. 845–848.

31. Красник У. Транслітерація прізвищ українських авторів у багатомовних каталогах бібліотеки / Уляна Красник. // Записки Львівської національної наукової бібліотеки України імені В. Стефаника. – 2012. – №4. – С. 263–275.

