ВІННИЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ

Факультет комп’ютерних систем та автоматики
Кафедра комп’ютерних систем управління
Спеціальність 151 Автоматизація та комп'ютерно-інтегровані технології
Освітньо-професійна програма Інтелектуальні комп’ютерні системи

ЗАТВЕРДЖУЮ

Завідувач кафедри КСУ
Дубовой В.М.

«___» _____ 2019 року

МАГІСТЕРСЬКА КВАЛІФІКАЦІЙНА РОБОТА

Інформаційна система ієрархічної категоризації науковців за відкритими даними з Google Scholar. Частина 1. Інтелектуальний аналіз даних

08-01.МКР.006.00.000

Студент групи 2АКІТ-18м Петричко М.В.

Керівник д.т.н., професор Штовба С.Д.

Рецензент к.т.н., доцент Софина О.Ю.

Вінниця ВНТУ – 2019

ВІННИЦЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
Факультет комп'ютерних систем та автоматики
Кафедра комп'ютерних систем управління
Освітньо-кваліфікаційний рівень магістр
Спеціальність 151 Автоматизація та комп'ютерно-інтегровані
Освітньо-професійна програма Інтелектуальні комп'ютерні системи

ЗАТВЕРДЖУЮ
Завідувач кафедри КСУ
Дубовой В.М.

«_2__» ____09_____ 2019року

Протокол № 1 засідання кафедри
КСУ від 2.09.2019р.

З А В Д А Н Н Я
НА МАГІСТЕРСЬКУ КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТУ

_____Петричку Миколі Володимировичу_____
(прізвище, ім'я, по батькові)

1. Тема магістерської кваліфікаційної роботи «Інформаційна система ієрархічної категоризації науковців за відкритими даними з Google Scholar. Частина 1. Інтелектуальний аналіз даних»
керівник магістерської кваліфікаційної роботи Штовба Сергій Дмитрович, д. т. н., професор _____
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)
затверджені наказом вищого навчального закладу від "02" 10 2019 року № 254

2. Строк подання студентом магістерської кваліфікаційної роботи 10.12. 2019 року
3. Вихідні дані до магістерської кваліфікаційної роботи мова інтересів науковця - англійська, підтримка ОС – Windows , Android; максимальний час завантаження – 5 с; авторизація користувачів – ні; максимальна кількість запитів до системи – до 1000 шт/с; мови графічного інтерфейсів – англійська.
4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити) вступ, аналіз систем підтримки наукових досліджень, задача категоризації та задача категоризації науковців, аналіз систем категоризації науковців, формування задач дослідження, аналіз існуючих робіпо категоризації науковців, запропонований метод вирішення поставленої задачі, обгрунтування вибору класифікації наук, модель схожості наукових спеціальностей, визначення схожості між науковцями, розробка програмного забезпечення, економічна частина.

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень) UML діаграма варіантів використання, UML-діаграма діяльності, UML-діаграма класів, ER-діаграма структури бази даних.

6. Консультанти розділів магістерської кваліфікаційної роботи

Розділ змістової частини роботи	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Економічний розділ	канд. економ. наук, доцент кафедри ЕПВМ Ратушняк О.Г.		
Розділ 1-3	д.т.н., професор кафедри КСУ Штовба С.Д.		

Календарний план

№ з/п	Назва етапів роботи	Строк виконання етапів роботи	Примітка
1	Вступ. Постановка задач дослідження	10.09.2019	
2	Розробка методу категоризації	20.10.2019	
3	Програмна реалізація	05.11.2019	
4	Економічний розділ	15.11.2019	
5	Апробація результатів дослідження	22.11.2019	
6	Публікації	01.12.2019	
7	Оформлення пояснювальної записки, графічного матеріалу і презентації	10.12.2019	
8	Захист МКР	12.12.2019	

Дата видачі завдання "_02__" __09___2019 року

Студент _____ _Петричко М.В.

Керівник магістерської кваліфікаційної роботи _____ _Штовба С.Д.

АНОТАЦІЯ

У роботі представлено метод багатоярликової ієрархічної класифікації науковців за їх інтересами в системі Google Scholar. Розроблений метод дає можливість автоматично віднести науковця до декількох релевантних наукових спеціальностей та галузей в системі ANZSRC. Першим науковим результатом роботи є модель оцінювання схожості наукових спеціальностей на основі статистики публікацій в системі Dimensions на основі індексу Жаккара. Другим науковим результатом роботи є удосконалений алгоритм категоризації науковців, який: а) окрім окремих інтересів враховуються ще і пари інтересів науковця,; б) під час редукції списку спеціальностей враховується їх схожість за запропонованою моделлю. На основі наукових результатів створена відповідна інформаційна система.

Ключові слова: Google Scholar, Dimensions, ANZSRC, наукометрія, профіль науковця, наукові інтереси, категоризація науковців, рейтингування науковців, інформаційна система, схожість наукових спеціальностей.

ABSTRACT

In the work, a method for multi-label hierarchical classification of researchers based on their interests in Google Scholar is presented. The method allows to automatically classify a researcher to a few research specialities and domains that fit for him the most in ANZSRC system. The first research result is the model to assess similarity of research specialities based on statistic of publications in Dimensions and Jaccard's index. The second research result is improved categorization method that: a) in addition to interests the pair of interests of a researcher is used; b) research specialities reduction uses specialities' similarity. Based on research results the information system was created.

Key words: Google Scholar, Dimensions, ANZSRC, scientometric, researcher's profile, research interests, researchers' categorization, researchers' ranking, information system, research specialities' similarity.

ЗМІСТ

ВСТУП

Останнім часом набуває поширення застосування штучного інтелекту у всіх галузях людської діяльності. Уже давно розроблені системи рекомендацій контенту, системи виявлення втручання, системи виявлення токсичного контенту в соціальних мереж тощо. Усі ці системи базуються на віднесенні того чи іншого об’єкту до одного з класів, тобто на класифікації. Останнім часом в задачах класифікації акценти змістили в бік обробки неструктурованої або слабкоструктурованої інформації, зокрема різноманітних природномовних текстів, які генеруються користувачами. Цей контент віддзеркалює уявлення користувача щодо того чи іншого явища, але це уявлення не формалізовано в межах деякої змістовної структури. Текстове вираження думок користувача значною мірою залежить від його персональних знань та лексикону. Відповідну одну і ту саму думку, люди висловлюють у різний спосіб. Така діяльність характерна для он-лайнових соціальних мереж.

Наукові спільноти теж об’єднані в різноманітні мережі. Найбільшою серед них є Google Scholar. В цій мережі у відкритому доступі є понад 50 тисяч профілів українських науковців. Науковець в профілі вказує свої інтереси, при чому робить він це на власний розсуд, обираючи слова у довільний спосіб. Google Scholar дозволяє здійснити пошук науковців за тим чи іншим інтересом. Але видачі формуються за буквальним співпадінням. Наприклад, видачі для “fuzzy set” та “fuzzy sets” будуть різними, не говорячи вже синонімічні інтереси типу “fuzzy evidence” та “fuzzy inference”. Також не враховує Google Scholar і сукупність інтересів користувача. Таким чином, пошукові та аналітичні сервіси за велетенським масивом профілів науковців в Google Scholar є досить примітивними.

Актуальною задачею є використання великих масивів відкритих даних Google Scholar для аналітичної діяльності. Для цього початкову слабкоструктуровану інформацію слід формалізувати у деякий спосіб, здійснити класифікацію науковців у деякій впорядкованій схемі. Подібним питанням займається бібліометрика

української науки – єдине джерело українських науковців, що класифіковані за галуззями науки. Проте в існуючій системі дана проблема вирішується суб'єктивно, тобто, задача класифікації науковця виконується людиною. Окрім цього відношення до галузі науки одиничне хоча науковець може бути обізнаним у декількох галузях науки. Враховуючи те, що науковці змінюють напрямки своєї діяльності такий підхід є неточним та затратним у плані людських ресурсів.

Мета дослідження. Метою роботи є розробка інформаційної системи ієрархічної категоризації науковців за відкритими даними з Google Scholar.

Об'єкт дослідження – інформаційні системи підтримки наукової діяльності.

Предмет дослідження - моделі та алгоритми ієрархічної категоризації науковців.

Методи дослідження. У дослідженнях використовувались методи теорії реляційних баз даних для формалізації моделей та операцій над даними науковців; методи прикладної статистики, комп'ютерної лінгвістики та інтелектуального аналізу даних для розробки методу категоризації науковців; методи об'єктно-орієнтованого програмування для розробки програмних складових інформаційної системи.

Наукова новизна отриманих результатів:

1. Вперше запропонована модель оцінювання схожості наукових спеціальностей на основі статистики публікацій в системі Dimensions, яка на відміну від моделей на основі відношення належностей пристатейно переліку джерел та цитувань до різних спеціальностей, використовує індекс Жакара за статями, що віднесені в системі Dimensions до двох спеціальностей.
2. Удосконалено алгоритм категоризації науковців за профілями в Google Scholar шляхом врахування взаємодії, а саме: а) окрім окремих інтересів враховуються ще і пари інтересів науковця; б) під час редукції списку спеціальностей враховується їх схожість за запропонованою моделлю.

Практичне значення одержаних результатів полягає у покращенні достовірності та зменшенні тривалості категоризації науковців. Окрім того, запропонована інформаційна система, на відміну від інших, категоризує науковців не за їх публікаціями, а за їх інтересами. Це дозволяє категоризувати молодих науковців, у яких мало або взагалі немає публікацій. Також виключається можливість помилкової категоризації у випадку публікацій у співавторстві, коли різним співавторам належать приципово різні наукові складові однієї статті. Результати категоризації за спеціальностями дозволяють автоматично категоризувати науковців і за галузями знань, тобто забезпечити ієрархічну категоризацію. Запропоновані моделі та алгоритми можна використовувати для створення інформаційних технологій ранжування науковців, підбору науковців з спільними інтересами для рецензування, опонування, виконання спільних проектів тощо, для моніторингу та визначення тенденцій щодо занятості науковців у різних напрямках тощо.

Апробація результатів роботи. Основні результати досліджень доповідалися на XLVII та XLVIII Науково-технічніх конференція факультету комп'ютерних систем і автоматики ВНТУ в 2018 р. та The Second International Workshop on Computer Modeling and Intelligent Systems, Запоріжжя, 2019 р.. Також прийнята доповідь на Digital Content & Smart Multimedia Workshop, який відбудеться в грудні 2019 р. у Львові.

Публікації. Результати дослідження опубліковано в 4 наукових статтях. Одна стаття у виданні із міжнародної бази Scopus [46], одна у фаховому журналі «Наукові праці Вінницького національного технічного університету» [51], та 2 статті у збірниках праць конференції [22, 47].

Дана робота виконана в межах держбютжетної теми ВНТУ 46-Д-388 «Ідентифікація прихованих залежностей в онлайнових соціальних мережах на основі методів нечіткої логіки та комп'ютерної лінгвістики», де автор працює виконавцем.

1 ОГЛЯД СТАНУ ПИТАННЯ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Інформаційні системи підтримки наукової діяльності як об’єкт дослідження

Останнім часом набувають популярності так звані системи підтримки наукової діяльності. Такі системи забезпечують дослідницькі організації та науковців інформацією та інструментами для покращення їхніх наукових можливостей, якостей та продуктивності [48]. До найпопулярніших з них відносять: Scopus, Web of Science (WoS), Google Scholar, ResearchGate, Dimensions та інші. Розглянемо детальніше деякі з них.

Уся наука в Scopus розбита на 27 галузей. В Scopus накопичено величезний обсяг наукових публікацій, кожна з яких категоризована за цими галузями. Scopus є великою базою анотацій та цитувань рецензованої літератури, зокрема журнальних статей, книг та матеріалів конференцій. Наповнення інформації здійснюється через журнали з відбором якісних. Усі журнали, що потрапляють до Scopus проходять відбір на якість.

На даний момент українські науковці широко застосовують систему Google Scholar. Це інструмент пошуку науковців створений компанією Google 18 листопада 2004 року. Головна мета – забезпечити простий спосіб пошуку наукової літератури будь-якої тематики. Google Scholar функціонує подібним чином до пошукового двигуна Google надаючи найбільш відповідні результати пошуку. Результатом пошуку у Google Scholar є мільйони наукових документів (анотацій, статей, тез, книг, статей з конференцій, патентів та ін.).

Dimensions – частково безкоштовна наукова база даних запущена компанією Digital Science у січні 2018 року. Містить у собі статі журналів та показники цитувань. Налічує понад 100 млн публікацій. Наука поділяється на 22 галузі та 154

спеціальності. Категоризація науковців здійснюється на основі категоризації статей. Категоризація статей здійснюється на основі аналізу ключових слів та анотації статті з використанням мтодів машиного навчання.

WoS є найбільш надійною, незалежною від видавництва, базою цитувань. Основний недолік є те, що статті враховуються переважно на англійській мові. Окрім цього у цій системі у більшій мірі представляють роботи з медицини, біології, психології, економіці, фізиці, астрономії. У меншій мірі – з математики, юриспруденції, політології і комп'ютерними наукам. Джерело інформації як і у випадку з Scopus – журнали. Категоризація науковців здійснюється за статтями.

ResearchGate є одночасно соціальною мережею і науковою базою документів. Основною відмінністю від попередніх систем є те, що джерелом інформації є самі науковці, саме вони завантажують наукові публікації у систему.

1.2 Задача категоризації науковців

Розглядається задача класифікації (рис. 2.1), тобто віднесення об'єкта з ознаками (атрибутами) $X = (x_1, x_2, ..., x_n)$ до одного із класів $\{l_1, l_2, ..., l_m\}$. З математичної точки зору класифікація – це відображення виду $X = (x_1, x_2, ..., x_n) \to y \in \{l_1, l_2, ..., l_m\}$. До класифікації зводяться різноманітні задачі прийняття рішень.

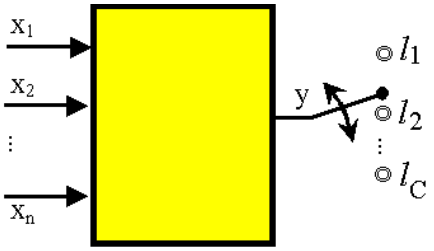


Рисунок 1.1 – Задача класифікації

Вхідні змінні можуть приймати значення з різних шкал. Найчастіше зустрічаються такі типи шкал даних [50]:

числова, наприклад, діапазон $[0, 24]$ для оцінювання середньої тривалості перебування в науковій мережі протягом доби;

категоріальна або номінальна, наприклад, множина {чоловіча, жіноча} для опису статі людини або множина {червоний, жовтий, зелений} для опису кольору. Частинним випадком категоріальної шкали є бінарна, коли атрибути приймають одне із двох можливих значень, наприклад, з множини {0, 1}, {ні, так}, {немає, є};

порядкова, наприклад, множина {незадовільно, задовільно, добре, відмінно} для оцінювання знань студента.

Багатоярликовий класифікатор (multi-label classifier) функціонує таким чином, що кожен об’єкт може бути віднесено не одного класу, а до кількох. Таку класифікацію також називають категоризацією. Наприклад, наукова стаття може бути віднесена до кількох наукових напрямків (рис.1.2).

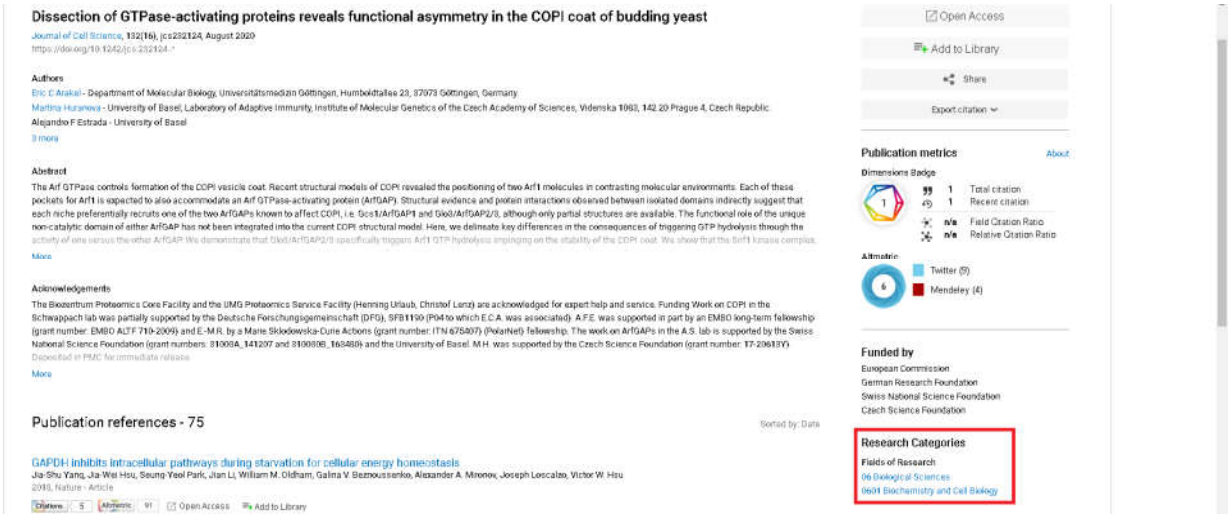


Рисунок 1.2 – Категоризація статті у Dimensions

Класифікація може бути детермінована та ймовірнісна. За детермінової класифікації рішення приймається за бінарної схеми – належить чи не належить. Тобто, вказується клас до якого віднесено об’єкт чи множина класів за категоризації. За ймовірносної класифікації вказується числова міра належності до кожного класу.

З швидким розвитком інформаційних технологій зростає зв’язність між учасниками різноманітних спільнот, в тому числі і наукових. Постає складна задача кластеризації учасників спільнот на групи з схожими інтересами. Стосовно наукових спільнот – це виявлення груп науковців, що працюють в схожих напрямках. Актуальність таких задач в Україні стрімко зросла в зв’язку із запровадженням в березні 2019 р. експерименту щодо захисту PhD-дисертацій в одноразових спецрадах. Однією із ключових умов формування одноразових спецрад є близькість наукового напрямку дисертації та наукових досліджень членів спецради. Виникає питання як формалізувати цю відстань, щоб автоматично верифікувати відповідність складу спецради науковому напрямку дисертації. Аналогічна задача виникає в рекомендаційних системах із пошуку партнерів для спільних наукових досліджень, пошуку друзів у наукових соціальних мережах тощо.

Категоризація науковців проводиться з декількох причин. По-перше, необхідно знати чим займається інститут, його конкретні напрямки розвитку. По-друге, категоризація науковців дозволяє виконати групування науковців за галузями або спеціальностям наук. Це дає можливість швидкого пошуку науковця, наприклад, для роботи над спільним проектом або опонентів для дисертації. По-третє, з’являється можливість визначити які саме галузі науки та спецальностиі популярні в даному інституті, університеті, регіоні, державі. Для держави це може бути корисним при формуванні державного бюджету таким чином, щоб підтримати розвиток тих галузей науки, що значно відстають від інших.

Як мотиваційний приклад розглянемо процес вибору рецензентів на конференцію для якої голова секції хоче обрати потенційних кандидатів для рецензування за темою секції «information extraction». При наявності системи пошук експертів голова секції може отримати список рекомендацій експертів для формування списку рецензентів по темі «information extraction». Без наявності єдиної бази науковців цю задачу вирішити досить складно та затратно у часі. Але при

наявності системи, у якій науковці категоризовані за галуззями наук задача зводиться до простого пошуку за ключовим словом.

На практиці пошук експерта вирішуються через особисту взаємодію та індивідуальні пошуки – рекомендації колег, запити до технічних груп або пошук кандидатів з даними навичками у база даних чи професійних онлайн мережах. Як результат, якщо соціальна мережа обмежена або пошукові групи не володіють наявними навичками для вирішення проблеми пошуки технічного співробітника можуть виявитись затратними у часі, ненадійним та виснажливим завданням без гарантій отримання бажаного результату.

Потенційне вирішення до проблеми знаходження необхідного співробітника це використання рекомендаційної системи. Категоризація науковців є однією із передумов (хоча це може бути не лише категоризація) до створення рекомендаційної системи науковців. Технології рекомендаційної системи надають автоматизований і надійний розв’язок спрощуючи задачу пошуку експертів та сприяючи роботі організації [3].

1.3 Аналіз систем категоризації науковців

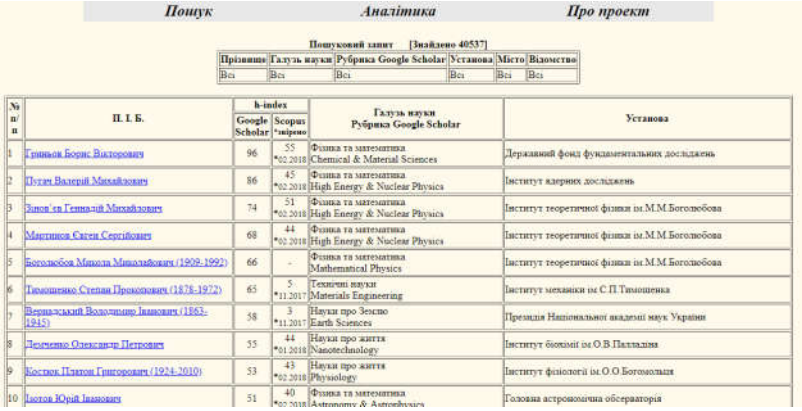
Прямим аналогом нашої системи є бібліометрика української науки [3]. Інформацію про науковців у даній системі сформовано на основі профілів Google Scholar. Типовий результат пошукового запиту у бібліометриці подано на рисунку 1.3. Окрім показників Google Scholar для кожного науковця вказано галузь науки. У бібліометриці класифікація науковця проводиться для таких галузей:

1) аграрні науки;
2) гуманітарні науки;
3) економіка;

4) інформатика;
5) медицина;
6) науки про життя;
7) науки про Землю;
8) педагогіка;
9) суспільні науки;
10) технічні науки;
11) фізика та математик;
12) хімія.

Поділ на науки у бібліометриці є дещо узагальненим. Насправді більшість з поданих галузей можуть поділятись на декілька підгалузей, що являють собою цілі науки. Головним недоліком є недостатній поділ на спеціальності науки, наприклад, галузь інформатика може поєднувати у собі Комп'ютерні науки, Штучний інтелект, Робототехніка, Нечіткі системи тощо. В результаті виникають непорозуміння при аналізі діяльності науковця. На відмінну від бібліометрики розроблена система дає можливість категоризації науковців до 22 галузей науки та приблизно до 154 підгалузей. Окрім цього є можливість ранжування науковців за належністю до галузі або підгалузі.

Рисунок 1.3 – Сайт української бібліометрики



Пошук Аналітика Про проект

Пошуковий запит [Знайдено 40537]

Прізвище	Галузь науки	Рубрика Google Scholar	Установа	Місто	Відомство
Всі	Всі	Всі	Всі	Всі	Всі

№ п/п	П. І. Б.	h-index Google Scholar	h-index Scopus *вирано	Галузь науки Рубрика Google Scholar	Установа
1	Гриньов Борис Вікторович	96	55 *02.2018	Фізика та математика Chemical & Material Sciences	Державний фонд фундаментальних досліджень
2	Пугач Валерій Михайлович	86	45 *02.2018	Фізика та математика High Energy & Nuclear Physics	Інститут ядерних досліджень
3	Зінов'єв Геннадій Михайлович	74	51 *02.2018	Фізика та математика High Energy & Nuclear Physics	Інститут теоретичної фізики ім.М.М.Боголюбова
4	Мартинов Євген Сергійович	68	44 *02.2018	Фізика та математика High Energy & Nuclear Physics	Інститут теоретичної фізики ім.М.М.Боголюбова
5	Боголюбов Микола Миколайович (1909-1992)	66	-	Фізика та математика Mathematical Physics	Інститут теоретичної фізики ім.М.М.Боголюбова
6	Тимошенко Степан Прокопович (1878-1972)	65	5 *11.2017	Технічні науки Materials Engineering	Інститут механіки ім.С.П.Тимошенка
7	Вернадський Володимир Іванович (1863-1945)	58	3 *11.2017	Науки про Землю Earth Sciences	Президія Національної академії наук України
8	Дем'яненко Олександр Петрович	55	44 *01.2018	Науки про життя Nanotechnology	Інститут біохімії ім.О.В.Палладіна
9	Костюк Платон Григорович (1924-2010)	53	43 *02.2018	Науки про життя Physiology	Інститут фізіології ім.О.О.Богомольця
10	Лютов Юрій Іванович	51	40 *02.2018	Фізика та математика Astronomy & Astrophysics	Головна астрономічна обсерваторія

Рисунок 1.4 – Список науковців в бібліометриці української науки

Окрім бібліометрики категоризація науковців може здійснюватися у системах підтримки наукової діяльності таких як Scopus, Dimensions, Google Scholar тощо. Більшість цих систем здійснюють категоризацію на основі статей науковця, хоча цей підхід має свої недоліки. По-перше, наукові публікації можуть писатись декількома науковцями з різним доменом знань і визначити, насправді кому яка частина роботи належить досить складно. По-друге, наукові статті можуть бути застарілими і неактуальними оскільки науковець за цей час міг змінити напрямок своєї діяльності. Також існує категоризація за спеціальностями дисертацій автора. Ще один вид категоризації - за журналами. Для наочності зобразимо існуючі варіанти здійснення категоризації та відповідних систем у яких вона застосовується у вигляді таблиці 1.2.

Таблиця 1.2 – Системи категоризацій

Категоризація	WoS	Dimensions	Scopus	OUCI
За статтями		**X**	**X**	
За журналами	**X**			**X**
За інтересами науковця				

Під час підготовки до виконання цієї роботи ми не знайшли інших робіт по даній темі, окрім згаданої вище бібліометрики української науки. Однак існують задачі, що можуть бути вирішені при наявності категоризованих науковців. Одна з

таких задач - рекомендація рецензента для написання рецензії. Роботи над цією тематикою мали місце у [1-4, 39-45].

Серед згаданих робіт цікавою є робота [39] у якій науковці моделюються використовуючи тематичне моделювання. За результатами ця робота є найближчою до нашої. У ній пропонується модель для рекомендації рецензента. Їхня модель має назву Автор-Дисципліна-Тема (Author-Subject-Topic, AST) (рис. 1.5). У ній інформація про дисципліни рецензентів вбудована для аналізу тематичних розподілів рецензованих робіт і публікації рецензентів. Їхня головна ідея полягає у тому, що експертиза науковця може бути представлена у ієрархічній формі де коренем є деяка макро концепція а листками є конкретні види контенту, що пов'язані з концепцією у корені.

Для перевірки моделі автори використали базу даних у галузі Information Systems and Management. За результатами дослідження цієї бази даних автори виявили, що науковці цієї галузі науки мають досить широкий діапазон наукових досліджень. Для прикладу, на рисунку 1.6, 1.7 подано діаграми розподілу тем двох науковців.

За своєю суттю результат роботи [39] є категоризацією науковців але в реальному житті науковець не може мати таку велику кількість інтересів у різних галузях науки. Така модель корисна для пошуку схожих науковців, пошуку рецензентів.

В роботі [23] запропонована рекомендаційна систему для спільних досліджень. Автори роботи поєднують тематичне моделювання, word2vec та так звану метрику mover distance для визначення подібності науковців на основі їх анотацій у статтях. За результатами система є успішною у визначенні існуючих спів-авторств з вибірки даних на зважаючи на те, що ця інформація не були використана при моделюванні. І при цьому пропонує валідні потенційні можливості для співпраці із споріднених галузей дослідження незалежно від попередньої співпраці. На рисунку 1.8 подано схематичне представлення цієї моделі. Подібність науковців визначається на основі

подібності їх публікацій, подібність публікацій визначається на основі подібності тематичного розподілу публікації, подібність тематичного розподілу визначається на основі подібності слів.

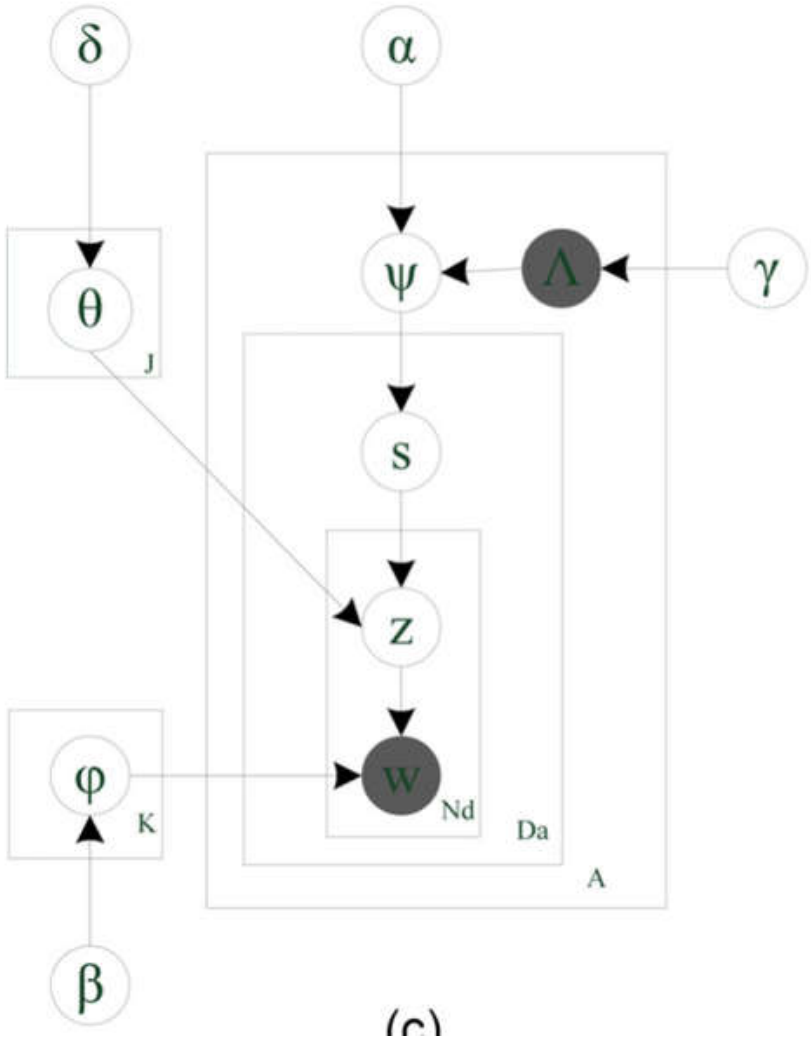


Рисунок 1.5 - Модель Автор-Дисципліна-Тема [39]

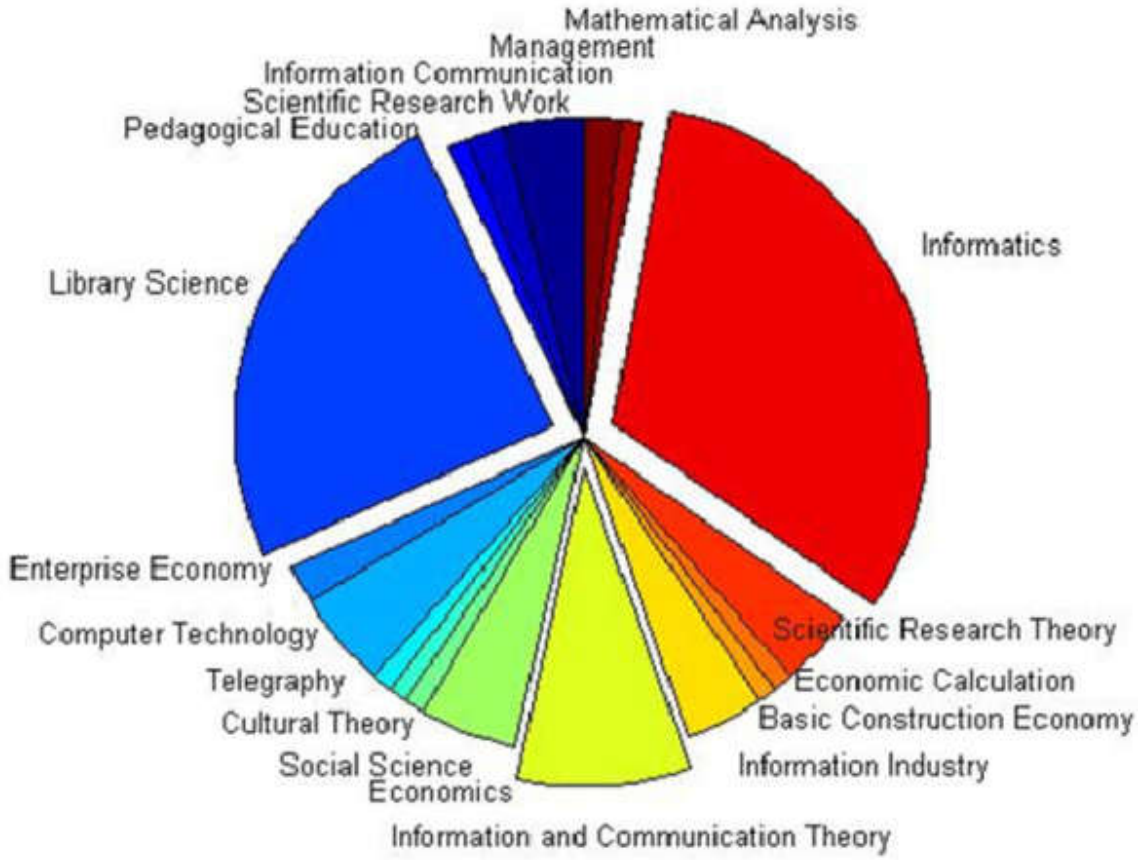


Рисунок 1.6 - Розподіл тем деякого науковця на основі моделі Автор-Дисципліна-Тема [39]

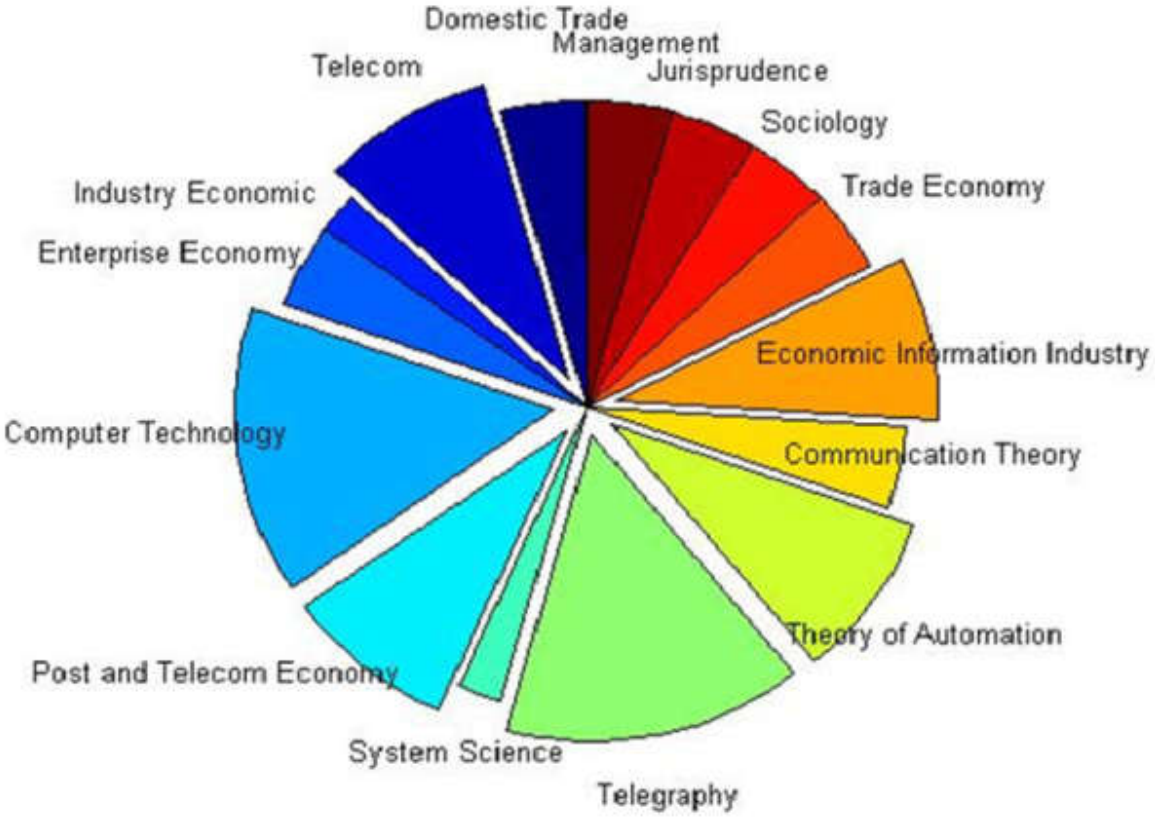


Рисунок 1.7 - Розподіл тем деякого науковця на основі моделі Автор-Дисципліна-Тема [39]

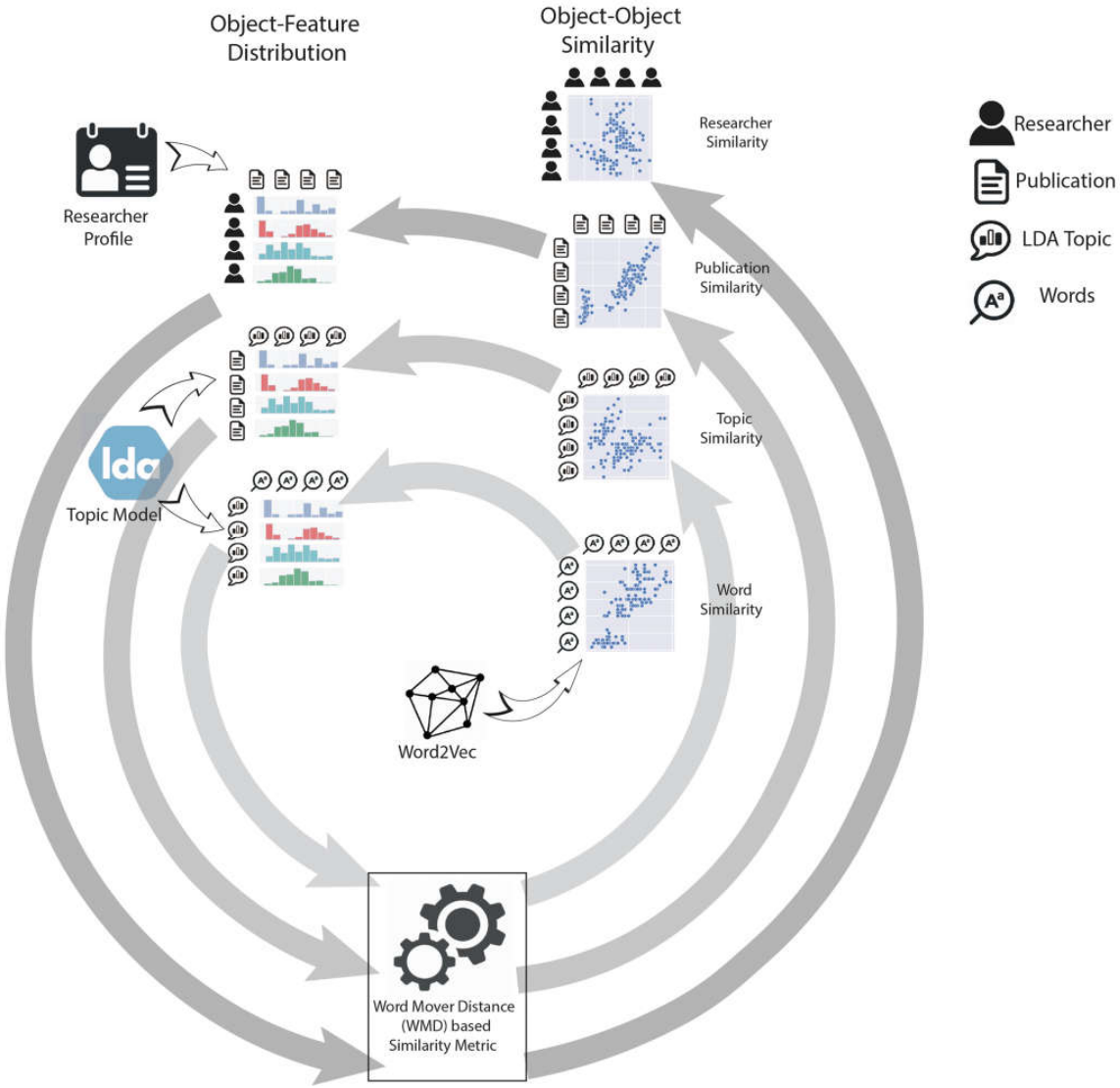


Рисунок 1.8 - Модель для знаходження подібності між науковцями [23]

1.4 Огляд існуючих підходів до задач моделювання природної мови

Задачі, що пов’язані з роботою над текстом за допомогою комп’ютера досліджує така галузь машинного навчання як обробка природної мови (Natural Language Processing, NLP). Підходи цієї галузі дозволяють аналізувати великі масиви тексту, робити їх семантичний аналіз, розпізнавати, синтезувати аудіо записи та реагувати на них, виконувати пошуки даних в контексті чогось, здійснювати рекомендацію різноманітного контенту тощо. Для представлення природної мови існують різноманітні підходи оскільки комп’ютер може розуміти лише 0 та 1. Базовою операцією NLP є представлення тексту у вигляді зрозумілому для комп’ютера.

Історія обробки природньої мови або Natural Language Processing (NLP) бере свій початок ще з 1950-х років хоча відомі роботи з більш ранніх періодів. На початку свого розвитку багато NLP систем були створені на основі так званих писаних від руки правил (Rule based): шляхом написання граматики або придумування евристичних правил для знаходження відмінностей між одиницями тексту або самим текстом [24]. З початку так званої «статистичної революції» у кінці 1980-х середині 1990-х більшість NLP досліджень почала значно покладатись на машинне навчання, Machine Learning (ML) [25]. ML парадигма для створення правил використовує статистичне виведення для формування правил обробки тексту базуючись на аналізі великої бази тексту, що зустрічається у реальному житті.

За весь час розвитку NLP було сформовано багато видів задач для вирішення. Деякі з цих задач мають прямі застосування на практиці інші ж використовуються як під задачі для вирішення більших задача. Хоча NLP задачі досить сильно перегукуються їх часто поділяють на категорії для зручності. Нище поданий список найпопулярніших NLP задач:

1) Задачі синтаксису:

1. Індукція граматики – навчання моделі для опису граматики у формальному вигляді [26];
2. Лематизація – приведення слова до його початкової форми;
3. Морфологічна сегментація – визначення класу морфем;
4. Часткове тегування (part-of-speech tagging) – виділення частин мови слів;
5. Парсинг – граматичний аналіз слів;
6. Розбиття речень – знаходження меж речення (ком, знаків пунктуації тощо);
7. Стемінг – знаходження основи слова;
8. Словесна сегментація – розділення тексту за словами;
9. Виділення термінології – знаходження релевантних тем у базі тексту;

2) Задачі семантики:

1. Лексична семантика – значень слів у контексті;
2. Розподілена семантика – способи семантичного представлення тексту;
3. Машинний переклад – автоматичний переклад тексту;
4. Розпізнавання сутностей – знаходження і визначення іменованих сутностей у тексті;
5. Генерування природньої мови – конвертування інформації з комп’ютера у природню мову;
6. Оптичне розпізнавання символів;
7. Відповідь на запитання;
8. Розпізнавання текстової поширеності – чи походить істинність одного тексту з іншого;
9. Виділення зв’язків між іменованими сутностями;
10. Тональний аналіз тексту – виявлення емоційної лексики;
11. Розпізнавання або сегментація тем – кластеризація елементів тексту на сегменти, що визначають теми;
12. Визначення значення двох-значних слів у тексті.

3) Задачі розпізнавання природної мови:
    1. Розпізнавання аудіо або природньої мови;
    2. Мовна сегментація – розбиття природньої мови на слова;
    3. Перетворення тексту у аудіо звук;

Як відомо, комп'ютер оперує лише числами і єдиний спосіб для нього зрозуміти природну мову – перетворити текст у числа. Методи для виконання даної операції розроблялись ще з початку виникнення NLP [27-34]. До найпопулярніших відносять:

1) Сумка слів (bag-of-words);
2) Частота терму - зворотна частота документу(TF-IDF – term frequency-inverse document frequency) [34];
3) Позитивна матрична факторизація [37];
4) Латентне розподілення Діріхле [35];
5) Латентний семантичний аналіз [36];
6) Тематичне моделювання;
7) Розподілене представлення слів і фраз та їхня композиційність (word2vec) [32];
8) Глобальні вектори для представлення слова (glove) [31];
9) Дво направлений енкодер представлення на основі перетворювачів (BERT) [27];
10) Глибокі контекстні представлення слова (ELMo) [28];
11) та багато інших.

Сумка слів

Даний метод представлення тексту у формі зрозумілій для комп'ютера є найпростішим і полягає у представленні відповідних елементів тексту їх індексами із словника. У випадку представлення слова це буде індекс слова у словнику, що створений із слів тексту, що опрацьовується. Цілий словник використовується для створення багато вимірного простору у якому кожний документ перетворюється у

вектор частоти терму (слова) через підрахунок появи термів у документі. Векторизація тексту є основою для багатьох сучасних алгоритмів обробки тексту [23]. Однак, використання термів як елементарних одиниць має свої недоліки. Для підкреслення обмеження представлення сумки слів використаємо приклад з [23] порівнюючи схожість між наступними трьома реченнями:

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

Перші два з цих речень мають досить схоже значення навіть враховуючи те, що різні слова були використані для опису ймовірно однієї і тієї ж події. Однак, звичайна евклідова відстань між частотними векторами термів не може цього визначити вона не враховує схожості між значення цих слів. Приведенні до нижнього регістру, перетворенні і лематизованні речення подано у таблиці 1. А косинус подібності між трьома векторами подано у таблиці 2. Відповідно до таблиці 2 речення 2 і 3 є ближчими ніж 1 і 2. Це показує важливість значення слів при порівнянні їх за допомогою комп’ютера та використання для різних NLP задач.

Таблиця 1.3 - Частота термів фраз з прикладу

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

Таблиця 1.4 - Матриця косинусів подібності серед векторизованих фраз

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

Word2Vec

Даний метод для представлення слова і фрази вважається одним із найкращих відомих на момент написання роботи. Суть цього методу [32] полягає у використанні так званої skip-gram [38] моделі - ефективного методу для навчання високоякісних векторних представлень слів з великої кількості не структурованої текстової інформації. Даний метод базується на нейронних мережах і на відміну від своїх аналогів більш ефективний у часі тому що не використовує повного перемноження матриць.

Цікавою особливістю цієї моделі є те, що просте сумування векторів може утворити інше представлення слова. Наприклад, vec(“Russia”)+ver(“river”) близьке до vec(“Volga River”), і vec(“Germany”)+vec(“capital”) близьке до vec(“Berlin”). Це свідчить про те що, певний рівень розуміння мови може бути отриманий використовуючи базові математичні операції над вектора представлення слів.

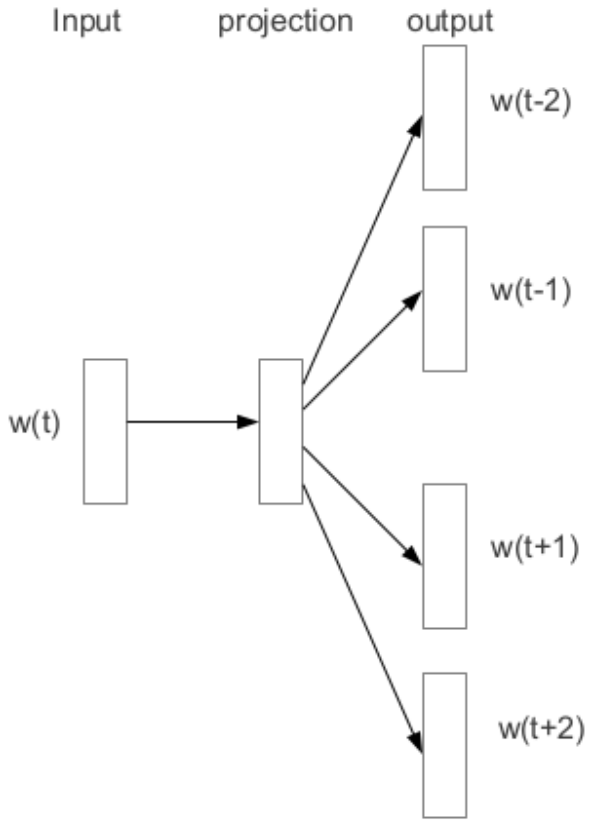


Рисунок 1.9 - Архітектура skip-gram моделі [38]. Ціль - навчити векторне представлення слова, що чудово передбачає сусідні слова.

Мета навчання skip-gram моделі - знайти представлення слова, що добре відображають навколишні слова у реченні або документі. У більш формальному визначенні, за наявності слів $w_0, w_1, .., w_T$, мета моделі це максимізація середньої логарифмічної ймовірності:

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

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

Рисунок 1.5 показує як word2vec модель може покращити проблему значення слів, що описана вище. З рисунку видно, що споріднені слова розташовані ближче у

просторі (тут використано метод зменшення просторовості для зручності представлення даних).

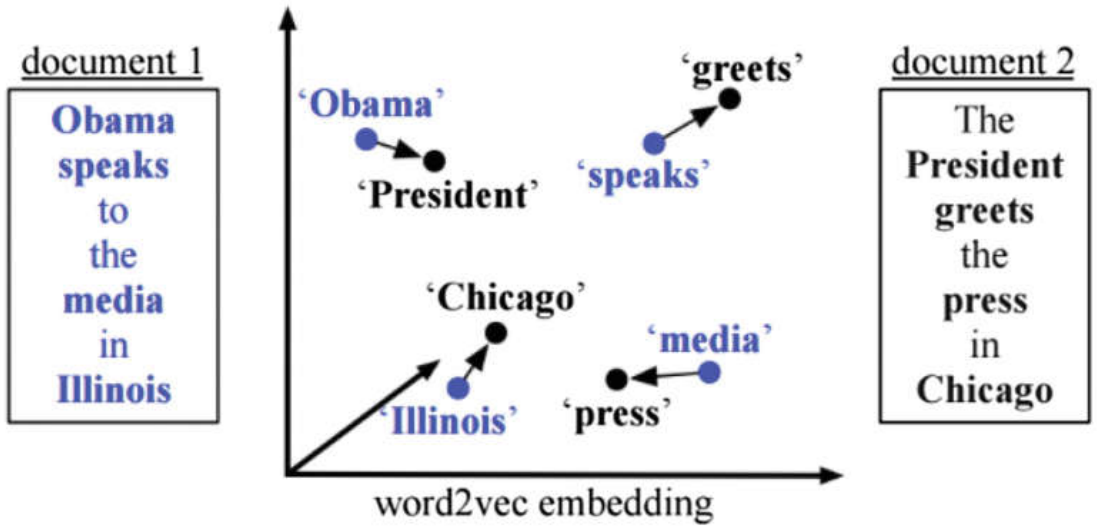


Рисунок 1.10 - Семантично схожі слова у просторі word2vec [23]

Glove

Ще один з найкращих методів представлення - глобальні вектори для представлення слів. На відміну від word2vec цей метод базується на факторизації матриць, що формуються з контексту слова [31]. Спочатку формується інформації спів-появи, тобто для кожного слова (рядки матриці) рахується частота його появи у деякому контексті (стовпчики) у великій базі тексту. Сам метод заснований на спостереженнях із таблиці 3 де зв'язок між словами ice і $steam$досліджується шляхом знаходження відношення їх спів-появ у деякому контексті k. Для слів k споріднених з ice але не з $steam$(наприклад, $k = solid$) очікується, що відношення їх спів-появ з k буде велике. І навпаки, для слів k пов'язаних з $steam$ але не з ice (наприклад, $k = gas$), співвідношення очікується малим. Саме на цьому спостереженні базується glove.

Таблиця 1.5 - Ймовірності спів-появи з обраним контекстом з бази тексту розміром 6 мільярдів токенів [31].

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

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

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

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

Тематичне моделювання

Тематичне моделювання (ТМ) це метод, що представляє документ у просторі меншої розмірності а ніж сумка слів оцінюючи вектор у просторі тем (topics) [35]. ТМ часто використовується як інструмент із видобування інформації з тексту (text-mining tool) для знаходження прихованих латентних структур у корпусі тексту, однак, розмірність і компоненти кожного вектору, представленні як “теми” (topics), не обов’язково корелюють із змістовою інтерпретацію яку людина може очікувати. Існуючі алгоритми тематичного моделювання згадані вище включають у себе: позитивна матрична факторизація [37]; Латентне розподілення Діріхле [35]; Латентний семантичний аналіз [36]. Найпопулярніший з них - Латентне розподілення Діріхле.

У тематичному моделюванні кожний документ представляється як ймовірнісний розподіл по усіх темах і кожна тема це ймовірнісний розподіл по усьому словнику корпусу.

1.5 Задачі дослідження

Мета комплексного дослідження полягає у розробці інформаційної системи ієрархічної категоризації науковців. Мета першої частини роботи полягає у створенні інформаційної системи для категоризації науковців за спеціальностями та галузями науки. Для досягнення мети необхідно виконати наступні завдання:

1) розробити модель представлення науковців у просторі галузей та спеціальностей за деякої специфікації наук;
2) дослідити різноманітні системи класифікації наук та обгрунтувати вибір найбільш придатної для вирішення задачі категоризації;
3) дослідити схожість наукових спеціальностей та розробити модель їх оцінювання;
4) покращити алгоритм категоризації науковців за спеціальностями яких враховує взаємодію як інтересів науковців, так і наукових спеціальностей ;
5) реалізувати програмно запропоновані моделі та алгоритми.

1.6 Висновки до розділу 1

У першому розділі розглянуто стан питання та сформовано постановку задачі. Проаналізовано різноманітні системи підтримки наукової діяльності.

Подано визначення категоризації та категоризації науковців і розглянуто для чого потрібна категоризація науковців.

Проаналізовано системи категоризації науковців, їх структуру, результат та місце нашої системи серед них. Проведено літературний огляд підходів до вирішення задачі категоризації науковців.

Сформовано мету роботи та завдання.

Проаналізовано сучасні підходи до вирішення задач обробки природньої мови.

Матеріали розділу опубліковано у працях автора [22, 47].

2 МЕТОД КАТЕГОРИЗАЦІЇ НАУКОВЦІВ

2.1 Ідея методу

Вважатимемо відомими:

$W = (w_1, w_2, ..., w_n)$ – перелік ключових слів та словосполучень тексту, якими описано наукові інтереси науковця в його профілі в Google Scholar ;

$T = (t_1, t_2, ..., t_m)$ – перелік можливих класів, тем – наукових спеціальностей за деякою класифікацією наук;

– множина розмічених текстів – множина публікацій кожна з яких віднесена до однієї або декількох тем (наук) T;

$R(D, T) \subset D \times T$ – відношення, яке описує належність публікацій до наук; $R(d_j, t_p) = 1$, якщо публікацію d_j віднесено до класу t_p, $j = \overline{1, k}$, $p = \overline{1, m}$.

Задача полягає у знаходженні тем (наук) з T, яким відповідає множина ключових слів (інтересів науковця) W. При цьому вказується не лише сам факт належності, але і ступінь належності. Таким чином на виході отримуємо нечітку множину $\widetilde{W}$ на підмножині універсальної множини T.

Пропонований метод вирішення цієї задачі базується на таких принципах:

Принцип статистичної підтримки – чим більша частка публікацій з певної науки містять аналізований інтерес, тим більша належність інтересу до цієї науки.

Принцип багатоярликовості – інтерес може належати до кількох наук.

Принцип фільтрації шумів – ігнорується незначний ступінь належності інтересу.

Принцип ігнорування стоп-слів – ігнорується інтерес, який зустрічається у дуже багатьох публікаціях.

Принцип солідарності – чим більше інтересів віднесено до однакової науки, тим сильніша можливість віднесення публікації до цієї науки.

Наведені принципи пропонується реалізувати наступним алгоритмом. В алгоритмі використовуються такі нові поняття:

p-та тематична колекція – підмножина з *D*, елементи якої віднесено до науки t_p, $p = \overline{1, m}$;

N_p – обсяг (кількість документів) *p*-ї тематичної колекції (колекції наук).

```
< Розрахувати обсяги тематичних колекцій N_1, N_2, ..., N_m >
< Counter = 0  – лічильник корисних інтересів >
for i=1:n
    < Знайти Q – кількість документів з D, в яких є
    інтересів w_i >
    If Q > Threshold_SW  continue; %ідентифікація стоп-слів
    end
    If Q < Threshold_Q  continue; %ідентифікація шуму
    end
    < Знайти кількість t_1(w_i), t_2(w_i), ..., t_m(w_i) документів
    тематичних колекцій з ключовим словом w_i >
    for p=1:m  % Занулюємо дуже малі видачі
        if    t_p(w_i) < Threshold_topic
        t_p(w_i) = 0
        end
    end
    If    max_{p=1,m} t_p(w_i) == 0  continue
    end
```

< Розраховуємо частоти входжень інтересу w_i в тематичні колекції $\gamma_p(w_i) = \frac{t_p(w_i)}{Q_p}$, $p = \overline{1,m}$ >

< Відбираємо мінімальну кількість найпопулярніших наук, сумарний внесок яких по γ не менше за $Tail\_1$. Номера невідібраних наук запишемо у вектор $Rejected$ >

< Занулюємо частоти входжень для непопулярних наук: $\gamma_{Rejected}(w_i) = 0$ >

< Нормуємо $\gamma(w_i)$ на 100% >

$Counter = Counter + 1$

$\mu_{Counter} = \gamma(w_i)$

end

If $Counter == 0$

Return ('Unsuccessful')

end

If $Counter > 1$

$\mu = \underset{u=1,Counter}{\text{mean}}(\mu_{Counter})$ %усереднюємо

< Відбираємо мінімальну кількість найпопулярніших наук, сумарний внесок яких по μ не менше за $Tail\_2$. Номера невідібраних наук запишемо у вектор $Rejected$ >

< Занулюємо частоти входжень непопулярних наук: $\mu_{Rejected} = 0$ >

< Нормуємо μ на 100% >

end

%Відрізаємо хвости – науки з дуже низькою належністю:

$Rejected = \text{find}(\mu(:) < Tail\_3)$

$\mu_{Rejected} = 0$

< Нормуємо μ на 100% >

End

В деяких задачах відомо, що W мають відповідати не більше ніж $Topic_{max}$ наук з T. Для цих випадків після головного циклу додаємо процедуру вилучення з розв'язку тем з рангами, більших за $Topic_{max}$.

2.2 Обгрунтування вибору системи класифікації наук та бази наукових публікацій

Для категоризації науковців необхідна система наук. В мережі є багато різноманітних класифікацій. Але при виборі ми враховуємо на лише систему класифікації а й базу наукових публікацій, яка буде використовуватись нашим алгоритмом для категоризації. Тому для нашого алгоритму необхідна база наукових публікацій, що виконує категоризацію публікацій за деякою категоризацією наук. І при цьому ця інформація повинна бути у публічному доступі.Серед таких систем найбільш популярними вважаються Scopus, Springer та Dimensions. Згідно з нашим алгоритмом такі системи також називаються базами розмічених документів. Вони надають інтерфейс для пошуку документів впорядкованих за галуззю/підгалуззю науки за ключовим словом.

Уся наука в Scopus розбита на 27 галузей (табл. 2.1). Система класифікації наукових досліджень у Скопусі містить 27 галузей. Виокремлення деяких великих галузей є сумнівним, і суперечить інших класифікаціям, наприклад, класифікації видавництва Wiley.

Таблиця 2.1 – Галузі Scopus

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

Springer – міжнародне видавництво, що займається публікацією книг, статей з журналів. Підтримує близько 3 тис. журналів та містить 250 тис. книг. Пропонує свої послуги для авторів, клієнтів та партнерів. Наука розбита на 24 галузі (рис. 2.1). Окрім галузей існує класифікація за підгалузями. Їх налічують близько 2 тис. З метою врахування лише значних підгалузей під час категоризації ми встановили

поріг входження підгалузі. Типовий пошук кількості публікацій за ключовим словом подано на рисунку 1.5. Результати пошуку знаходяться у блоках «Discipline», «Subdiscipline». В Springer наукові галузі погано впорядковані – містять як синоніми, так і занадто дрібні екземпляри. Більш цікавим є підгалузі (subdisciplines) Шпрингера, які дозволяють достатньо вузько категоризувати дослідження. Усього таких підгалузей майже 1600. Як і у випадку галузей, деякі підгалузі мають крихітне наповнення. Є також і синонімія. Нами встановлено, що 95% наповнення припадає на 595 підгалузей.

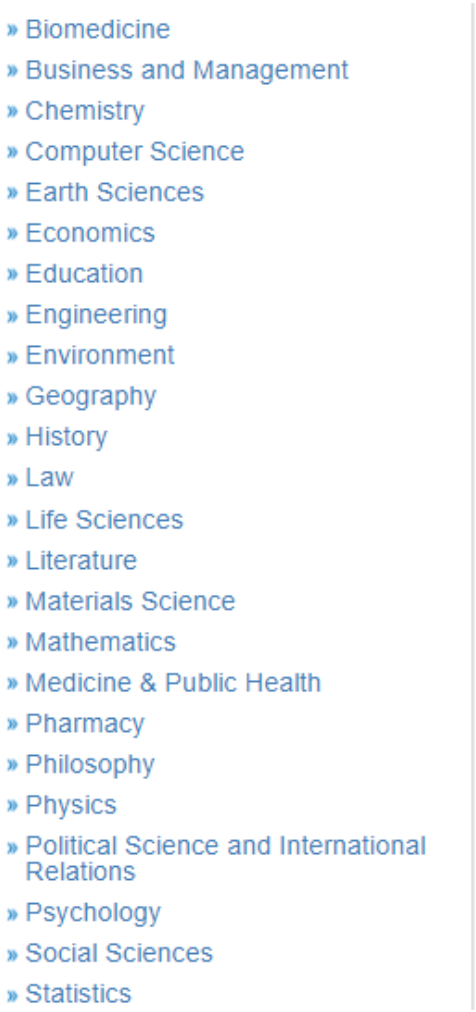


Рисунок 2.1 – Галузі науки у Springer

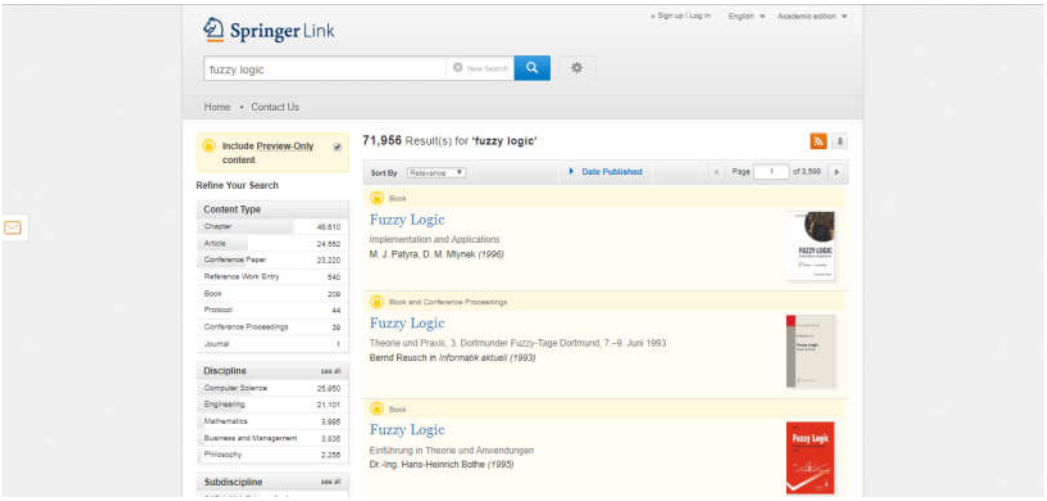


Рисунок 2.2 – Пошук кількості публікацій за ключовим словом у Springer

Dimensions містить у собі статі журналів та показники цитувань. Налічує близько 100 тис. публікацій. Наука поділяється на 22 галузі та 154 підгалузі. Типовий приклад пошуку кількості публікацій за ключовим словом подано на рисунку 2.3. Результат пошуку знаходиться у блоці «Fields of Research». Результат пошуку Dimensions містить публікації у яких згадано пошукове слово, кількість публікацій по кожній науці та спеціальності науки. Класифікація наук визначена у системі ANZSRC (Australian and New Zealand Standard Research Classification), що сформована Австралійським Бюро Статистики (Australian Bureau of Statistics) для оцінювання і аналізу наукових досліджень проведених у Австралії та Новій Зеландії.

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

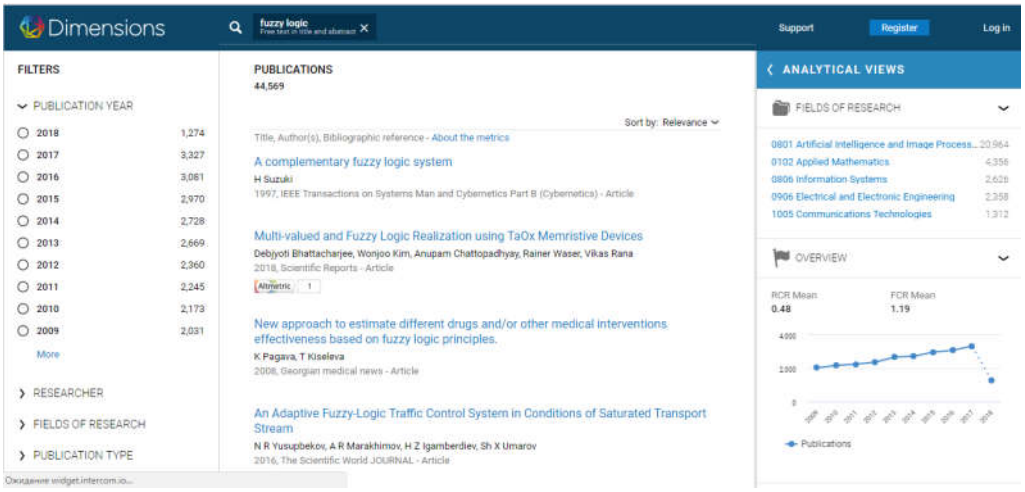


Рисунок 2.3 – Пошук публікацій за ключовим словом у науковій базі Dimensions

2.3 Алгоритм категоризації науковців в системі ANZSRC

Згідно з пропонованим алгоритмом для категоризації науковців необхідна так звана множина розмічених текстів D. У якості такої множини ми використовуємо множину публікацій науковців по усьому світу.

Галузі та підгалузі візьмемо з австарлійсько-новозеленської класифікації наук – Australian and New Zealand Standard Research Classification ANZSRC. Ця класифікація використовується в новій наукометричній системі Dimensions, яка через парсинг дозволяє реалізувати аналог перетворення словниково-вільних ключових слів в базисний набір ключових слів, а також в семантичні групи. Відповідно, ми матимемо можливість протестувати наші ідеї на великому обсязі реальних текстів.

Система ANZSRC має таку структуру:

01 Mathematical Sciences

0101 Pure Mathematics

0102 Applied Mathematics

- 0103 Numerical and Computational Mathematics
- 0104 Statistics
- 0105 Mathematical Physics

02 Physical Sciences
- 0201 Astronomical and Space Sciences
- 0202 Atomic, Molecular, Nuclear, Particle and Plasma Physics
- 0203 Classical Physics
- 0204 Condensed Matter Physics
- 0205 Optical Physics
- 0206 Quantum Physics
- 0299 Other Physical Sciences

03 Chemical Sciences
- 0301 Analytical Chemistry
- 0302 Inorganic Chemistry
- 0303 Macromolecular and Materials Chemistry
- 0304 Medicinal and Biomolecular Chemistry
- 0305 Organic Chemistry
- 0306 Physical Chemistry (incl. Structural)
- 0307 Theoretical and Computational Chemistry
- 0399 Other Chemical Sciences

04 Earth Sciences
- 0401 Atmospheric Sciences
- 0402 Geochemistry
- 0403 Geology
- 0404 Geophysics
- 0405 Oceanography
- 0406 Physical Geography and Environmental Geoscience
- 0499 Other Earth Sciences

05 Environmental Sciences

- 0501 Ecological Applications
- 0502 Environmental Science and Management
- 0503 Soil Sciences
- 0599 Other Environmental Sciences

06 Biological Sciences
- 0601 Biochemistry and Cell Biology
- 0602 Ecology
- 0603 Evolutionary Biology
- 0604 Genetics
- 0605 Microbiology
- 0606 Physiology
- 0607 Plant Biology
- 0608 Zoology
- 0699 Other Biological Sciences

07 Agricultural and Veterinary Sciences
- 0701 Agriculture, Land and Farm Management
- 0702 Animal Production
- 0703 Crop and Pasture Production
- 0704 Fisheries Sciences
- 0705 Forestry Sciences
- 0706 Horticultural Production
- 0707 Veterinary Sciences
- 0799 Other Agricultural and Veterinary Sciences

08 Information and Computing Sciences
- 0801 Artificial Intelligence and Image Processing
- 0802 Computation Theory and Mathematics
- 0803 Computer Software
- 0804 Data Format
- 0805 Distributed Computing

- 0806 Information Systems
- 0807 Library and Information Studies
- 0899 Other Information and Computing Sciences

09 Engineering

- 0901 Aerospace Engineering
- 0902 Automotive Engineering
- 0903 Biomedical Engineering
- 0904 Chemical Engineering
- 0905 Civil Engineering
- 0906 Electrical and Electronic Engineering
- 0907 Environmental Engineering
- 0908 Food Sciences
- 0909 Geomatic Engineering
- 0910 Manufacturing Engineering
- 0911 Maritime Engineering
- 0912 Materials Engineering
- 0913 Mechanical Engineering
- 0914 Resources Engineering and Extractive Metallurgy
- 0915 Interdisciplinary Engineering
- 0999 Other Engineering

10 Technology

- 1001 Agricultural Biotechnology
- 1002 Environmental Biotechnology
- 1003 Industrial Biotechnology
- 1004 Medical Biotechnology
- 1005 Communications Technologies
- 1006 Computer Hardware
- 1007 Nanotechnology

1099 Other Technology

11 Medical and Health Sciences

1101 Medical Biochemistry and Metabolomics

1102 Cardiorespiratory Medicine and Haematology

1103 Clinical Sciences

1104 Complementary and Alternative Medicine

1105 Dentistry

1106 Human Movement and Sports Science

1107 Immunology

1108 Medical Microbiology

1109 Neurosciences

1110 Nursing

1111 Nutrition and Dietetics

1112 Oncology and Carcinogenesis

1113 Ophthalmology and Optometry

1114 Paediatrics and Reproductive Medicine

1115 Pharmacology and Pharmaceutical Sciences

1116 Medical Physiology

1117 Public Health and Health Services

1199 Other Medical and Health Sciences

12 Built Environment and Design

1201 Architecture

1202 Building

1203 Design Practice and Management

1204 Engineering Design

1205 Urban and Regional Planning

1299 Other Built Environment and Design

13 Education

1301 Education Systems

- 1302 Curriculum and Pedagogy
- 1303 Specialist Studies In Education
- 1399 Other Education

14 Economics
- 1401 Economic Theory
- 1402 Applied Economics
- 1403 Econometrics
- 1499 Other Economics

15 Commerce, Management, Tourism and Services
- 1501 Accounting, Auditing and Accountability
- 1502 Banking, Finance and Investment
- 1503 Business and Management
- 1504 Commercial Services
- 1505 Marketing
- 1506 Tourism
- 1507 Transportation and Freight Services

16 Studies in Human Society
- 1601 Anthropology
- 1602 Criminology
- 1603 Demography
- 1604 Human Geography
- 1605 Policy and Administration
- 1606 Political Science
- 1607 Social Work
- 1608 Sociology
- 1699 Other Studies In Human Society

17 Psychology and Cognitive Sciences
- 1701 Psychology
- 1702 Cognitive Sciences

1799 Other Psychology and Cognitive Sciences

18 Law and Legal Studies
1801 Law

1899 Other Law and Legal Studies

19 Studies in Creative Arts and Writing
1901 Art Theory and Criticism

1902 Film, Television and Digital Media

1903 Journalism and Professional Writing

1904 Performing Arts and Creative Writing

1905 Visual Arts and Crafts

1999 Other Studies In Creative Arts and Writing

20 Language, Communication and Culture
2001 Communication and Media Studies

2002 Cultural Studies

2003 Language Studies

2004 Linguistics

2005 Literary Studies

2099 Other Language, Communication and Culture

21 History and Archaeology
2101 Archaeology

2102 Curatorial and Related Studies

2103 Historical Studies

2199 Other History and Archaeology

22 Philosophy and Religious Studies
2201 Applied Ethics

2202 History and Philosophy of Specific Fields

2203 Philosophy

2204 Religion and Religious Studies

2299 Other Philosophy and Religious Studies

2.4 Модель схожості наукових спеціальностей

Подібність науковців можна визначити через агрегування перетину за кожною спеціальністю. При цьому не враховується внесок споріднених спеціальностей, наприклад, геології та геохімії. Окрім цього при категоризації на етапі нехтування непопулярних наук не враховується підмножина наук, що є дуже спорідненими. Якщо таких наук багато то має сенс перерахувати вплив науки, що нехтується на споріднені науки, щоб збільшити ймовірність присутності споріднених наук на завершальному етапі алгоритму. Для визначення подібності можна застосовувати підходи з екології та ботаніки, наприклад, коефіцієнти Жакарда та Чекановського. Але вони враховують подібність за бінарним принципом. Метою нашого дослідження, є чисельна оцінка подібності споріднених спеціальностей. Оцінку проведено для переліку спеціальностей ANZSRC з використанням інформаційної системи Dimensions. Усе, що описано у цьому підрозділі базується на роботах автора [46, 47, 49].

Ми пропонуємо метод визначення подібності наукових спеціальностей, що базується на кількості публікацій які віднесені до цих спеціальностей. Для визначення коефіцієнту подібності введемо наступні позначення:

$S1$ – перша спеціальність;

$S2$ – друга спеціальність;

$k1$ – кількість документів по першій спеціальності;

$k2$– кількість документів по другій спеціальності;

C – кількість документів з галузями 1 та 2.

Тоді подібність цих двох галузей визначатиметься як:

$$similarity(S1, S2) = \frac{C}{k1+k2-C} \tag{2.1}$$

Значення $similarity(S1, S2)$ знаходиться в діапазоні $[0,1]$, де 0 означає відсутність подібності, 1 – ідентичність. Дану формулу схожості називають метрикою за Жаккаром.

Для усіх пар спеціальностей можна сформувати матрицю коефіцієнтів подібності:

$$similarity(S_i, S_j) = \frac{C_{ij}}{k_i + k_j - C_{ij}}, \quad (2.2)$$

де $i = \overline{1,n}$, $j = \overline{1,n}$, n – кількість спеціальностей; C_{ij} – кількість документів, які відносяться одночасно до i-ї та j-ї спеціальності; k_i – кількість документів, що віднесені до i-ї спеціальності; k_j - кількість документів, що віднесені до j-ї спеціальності.

Розглянемо приклад визначення коефіцієнту подібності використовуючи систему Dimensions. Візьмемо дві спеціальності Geology та Geochemistry. Дана система надає статистику кількості публікацій по кожній спеціальності науки. Для Geology та Geochemistry з відповідними кількостями публікацій 185265 та 60146 та кількістю публікацій одначсно з Geology та Geochemistry коефіцієнт подібності дорівнюватиме: $similarity(\text{Geology}, \text{Geochemistry}) = \frac{25994}{185265+60146-25994} = 0.1184$.

Маючи таку статистику можна визначити подібність між будь-якими двома спеціальностями у даній системі наук. Дослідження для схожості спеціальностей проводились для двох періодів 2000-2009 роки та 2010-2019 роки. Розподіл пар подібностей науковий спеціальностей на основі метрики Жаккара подано на рисунку 2.4. Більшість наукових спеціальностей мають нульову подібність. При порівнянні двох періодів виявилось, що перші пари 2000-2009 періоду мають більші коефіцієнти подібності, а ніж другому періоді. Але розподіл для 2010-2019 періоду має більше ненульових пар подібних спеціальностей. Найбільш схожі спеціальності за 2010-2019 період подано на рисунку 2.5. Найбільш схожі спеціальності з'являються у сфері освіти, планування міста та логістиці, мистецтві, комерційні науки та туризм тощо.

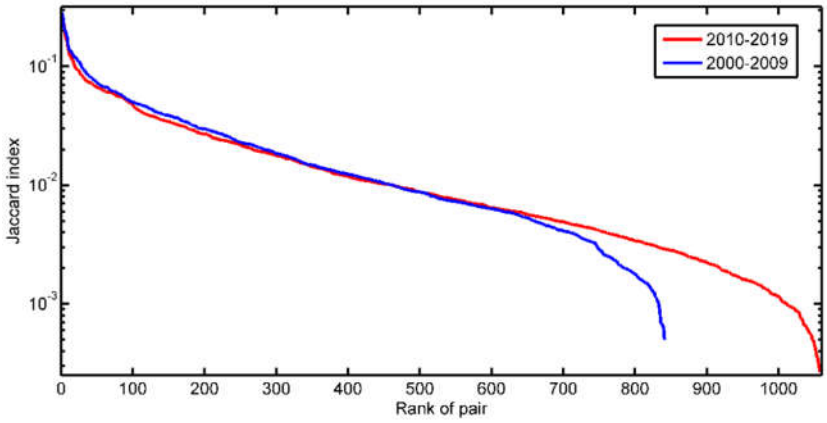


Рисунок 2.4 – Розподіл подібності пар наукових спеціальностей

Рисунок 2.5 – Найбільш схожі наукові спеціальності за 2010-2019 період

Одне із застосувань коефіцієнту подібності спеціальностей є пошук найближчих науковців. Припустимо, що у нас є база науковців, що віднесені з деяким ступенем належності до деякої спеціальності/спеціальностей із заданої системи наук, тоді найпростіший коефіцієнт подібності науковців має вигляд:

$$similarity(r1, r2) = sum(\min(A, B)), \quad (2.3)$$

де $r1, r2$ – деякі науковці, A та B – вектори ступенів належності до спеціальності/спеціальностей відповідних науковців. Така міра подібності називається ще мірою Чекановського. Вона має декілька недоліків. Один з них це те, що вона не враховує споріднені спеціальності. Очевидно, що якщо науковці віднесені до споріднених спеціальностей за даною ознакою подібності подібність дорівнюватиме нулю, оскільки їх перетин (мінімум) це нуль. Дану ознаку можна вдосконалити записавши:

$$similarity(r1, r2) = \sum_{j=0}^{n} \min(A_j, B_j) + \sum_{i=0}^{n} \sum_{j=0}^{n} similarity(s_i, s_j) * \min(A_i, B_j), \quad (2.4)$$

де $i = \overline{1, n}$, $j = \overline{1, n}$, n – кількість спеціальностей; A_i – ступінь належності науковця $r1$ до i-ї спеціальності, B_j – ступінь належності науковця $r2$ до j – спеціальності; s_i – i-та спеціальність, s_j – j-та спеціальність. Таким чином, за рахунок другої складової ознаки подібності два науковці з спорідненими спеціальностями завжди матимуть деяке значення подібності.

Розглянемо приклад знаходження подібності двох науковців. Для цього необхідно мати науковців, що віднесені до певної спеціальності/спеціальностей науки. Використаємо науковців з нашої системи категоризації науковців [3]. Для прикладу візьмемо двох науковців: Штовбу С.Д. та Бісікала О.В. (рис. 2.6). Порівняємо їхню схожість за подібністю спеціальностей науки. Розрахунки зведемо в табл. 2.2. Таким чином на основі подібності спеціальностей науки нам вдалось збільшити подібність науковців зі спорідненими спеціальностями науки.

Таблиця 2.2 – До розрахунку подібності наукових напрямків двох науковців

Науковець	Спеціальності	S1-S3	S1-S3	S2-S4	S3-S4	S1-S4	C1	C2
Бісікало О.В.	S1 - 0.396 S2 – 0.316 S3 – 0.286	0,0037184724	0,00383 6850	0,00269 967	0,00086 764	0,05908 7624	0.3 96	0,4179 330
Штовба С.Д.	S1 – 0.508 S4 – 0.491							

Розшифрування до таблиці:
S1 - Artificial Intelligence and Image Processing;

S2 – Linguistics;

S3 - Cognitive Sciences;

S4 - Information Systems;

C1 – Подібність за першою метрикою;

C2 – Подібність за метрикою схожості спеціальностей.

Рисунок 2.6 – Науковці для визначення подібності

2.5 Контрольні приклади категоризації

Проілюструємо роботу методу на прикладі категоризації учасників наукової мережі Google Scholar. Фрагменти профілів науковців для категоризації наведено на рис.2.7. На ньому вказано наукові інтереси, які ототожнимо з ключовими словами, що характеризують діяльність науковця. Слова, якими в профілі Google Scholar описано інтереси, науковці обирають на власний розсуд, без прив'язки до якогось словника. З іншого боку є класифікація наук, з прийнятими назвами тематик. Відповідно, ключові слова з профілю можна ототожнити з W, а галузі наук з деякої формальної класифікації через T. Тоді, виникає задача співставлення W з елементами множини T. Саме для вирішення цієї задачі і розроблено метод у цьому розділі.

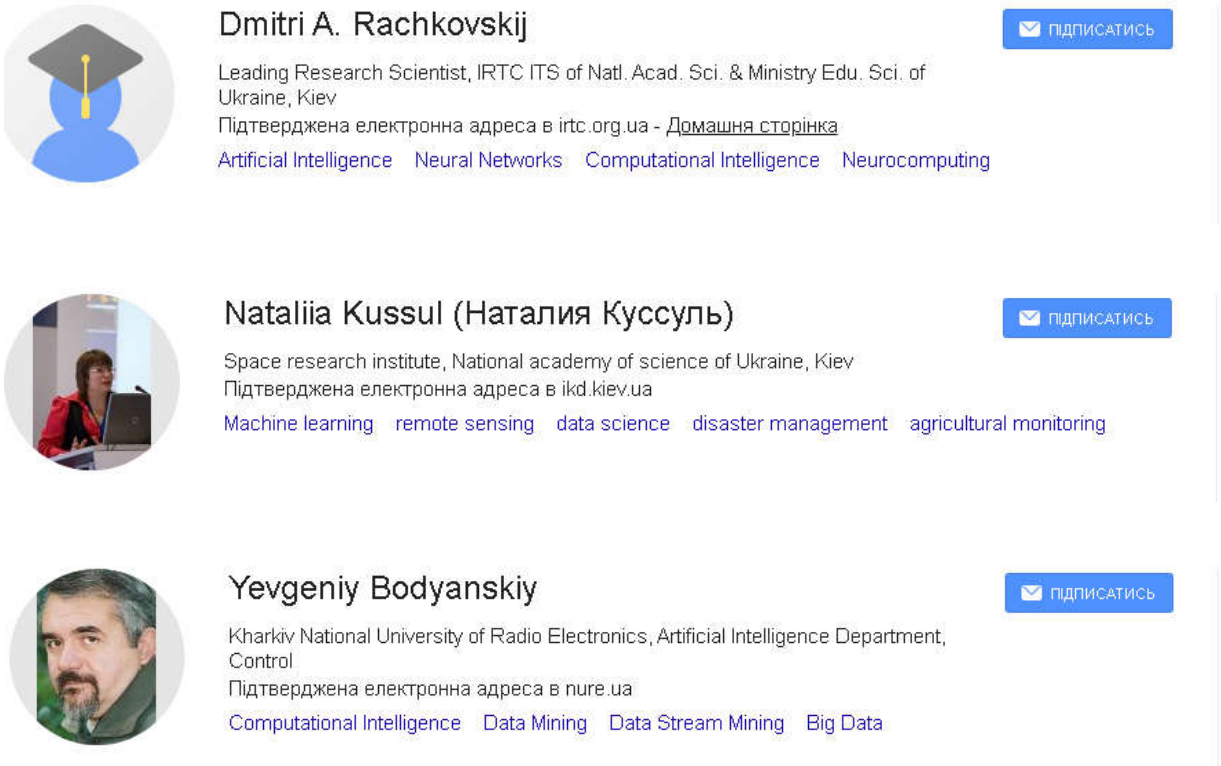


Рисунок 2.7 – Тестові профілі для категоризації

Основним джерелом даних для категоризації у нашій системі є інтереси науковця. Для науковців вище вказані наступні інтереси: *Artificial Intelligence, Neural Networks, Computational Intelligence, Neurocomputing, Machine Learning, remote sensing, data science, disaster management, agricultural monitoring, Data Mining, Data Stream Mining, Big Data*. Найперший етап категоризації – збір кількості появи відповідних інтересів у документах відповідних спеціальностях науки. Результат для вказаних інтересів подано на рисунку 2.8. Для спрощення парні запити не показано на рисунку. З рисунка видно, що ключове слово *"artificial intelligence"* має однозначне співпадіння з спеціальністю науки. Це зумовлено частиною нашого алгоритму у якій він обробляє ключові слова на предмет синонімів із спеціальностями.

Наступний етап полягає у відкиданні непопулярних тематик (спеціальностей науки). Тобто, у розподілі нехтуються значення тих тем, належність до яких незначна. Не значимість належності визначається кумулятивною сумою, що повинна складати 2/3 усіх рівнів належності. Після цього якщо у науковця більше одного інтересу виконується їх усереднення результат якого показано на рисунку 2.9.

Після усереднення розподілів інтересів знову виконується відкидання непопулярних тематик за тією ж схемою, що й для розподілу одного інтересу. Результат цього відкидання подано на рисунку 2.10. Результуючий розподіл може вийте перенаповненим через загальність інтересу. Через це введено ще один рівень фільтрації. За ним мінімальний ступінь належності до тематики становить 15%. Якщо тематик більше 6 то залишаємо лише 6 і після цього виконуємо наступні дії: видаляємо 6 тему і перераховуємо її внесок на інші 5 використовуючи їхню схожість; повторюємо ці дії для 5 тематики; якщо ступінь належності 4 тематики менше 15% виконуємо ці ж дії і для неї; рекурсивно перевіряємо для 3 і т.д. Мотивація для використання перерахунку ступенів належності на основі схожості спеціальностей пояснюється тим, що якщо у розподілі багато споріднених

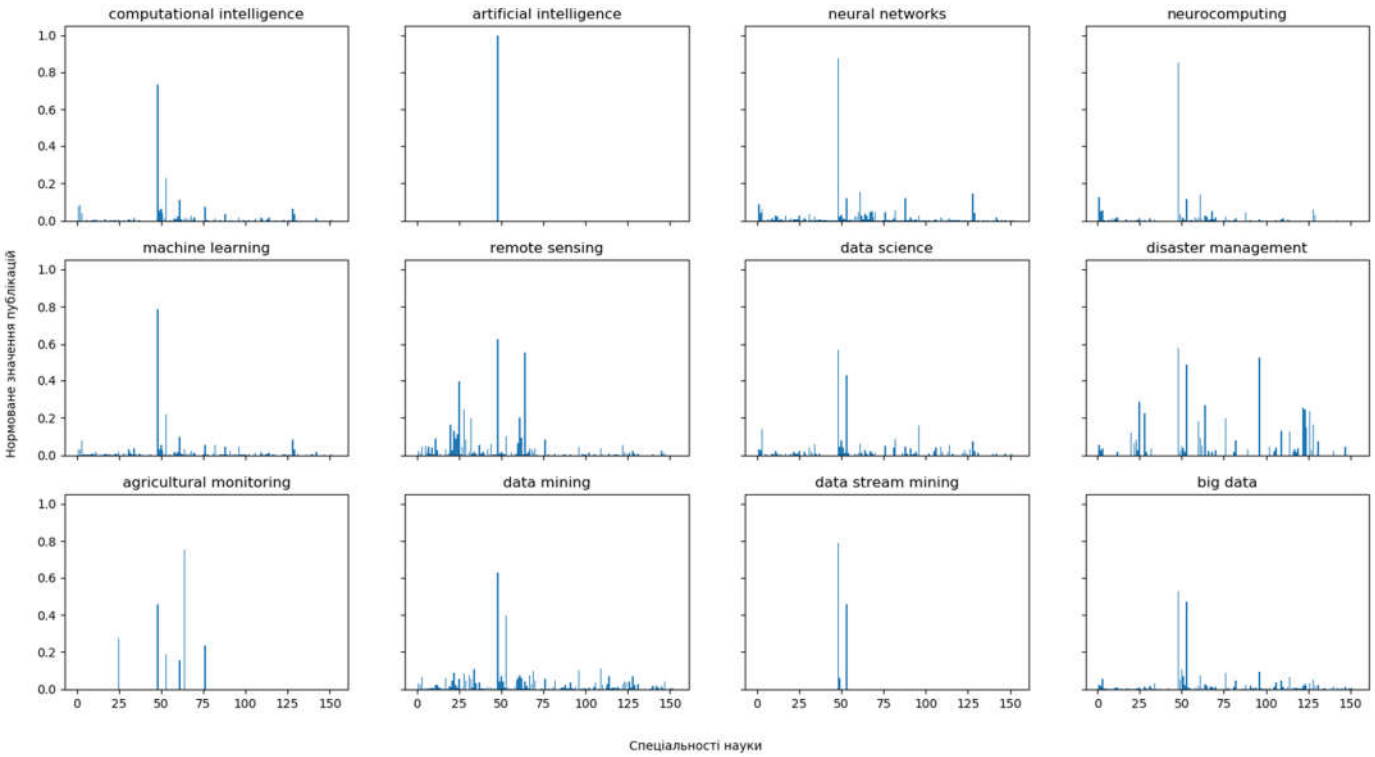


Рисунок 2.8 – Розподіл інтересів за спеціальностями науки

спеціальностей то має сенс їхній загальний внесок на результат категоризації. Формально перерахунок визначений у вигляді формули:

$$\bigvee_{i=0}^{n-1} A_i = A_i + A_i * similarity(S_i, S_n), \quad (2.5)$$

де A – множина ступенів належності, S – множина тематик. Множини відсортовані за спаданням ступенів належності. За рахунок перерахунку ступенів належності може змінитись порядок тематик і це може вплинути на наступні етапи відкидання непопулярних тематик. Результати за описаним типом фільтрації подано на рисунку 2.10. Даний етам фільтрації є кінцевим і те що зображено на рисунку 2.11 є результатом категоризаціі тестових науковців.

На даний момент ми працюємо над визначенням точності даної моделі. Точність даної моделі може бути визначена на основі подібності між науковцями, що описана у наступному підрозділі. Але у нас поки що немає еталонних даних для перевірки адекватності моделі. Більшість систем підтримки наукової діяльності категоризують науковців на основі їх статей, що є не придатним для порівняння з нашою моделлю. Для прикладу проведемо порівняння між категоризацією науковця у Dimensions і нашою. Результат категоризації Dimensions за статтями про нормуємо по сумі усіх статей науковця та залишимо лише 2/3 найпопулярніших. Результат категоризацій подано на рисунках 2.12-2.14. Згідно з рисунками більшість спеціальностей категоризації співпадають хоча є й такі яких немає у нашій категоризації. Схожість можна покращити врахувавши подібність між спеціальностями різних категоризацій.

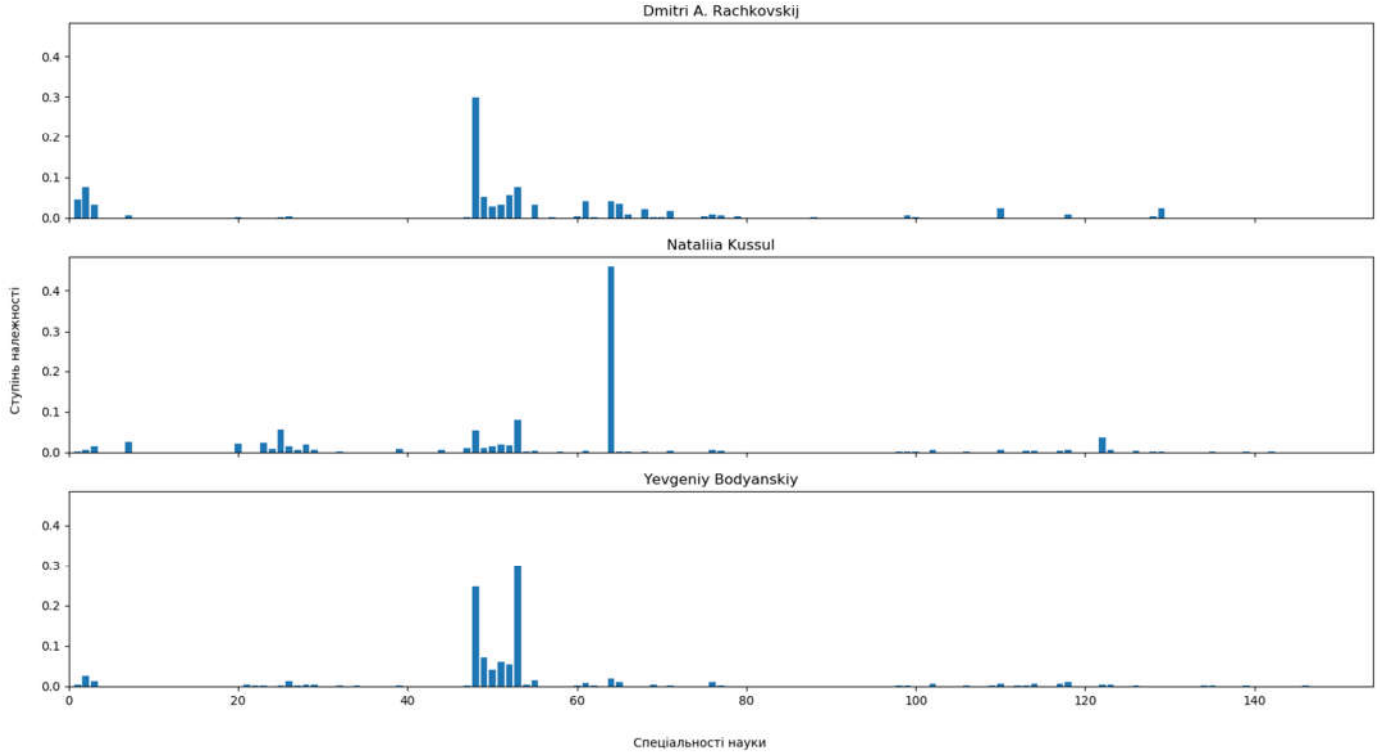


Рисунок 2.9 – Розподіл інтересів науковців після усереднення

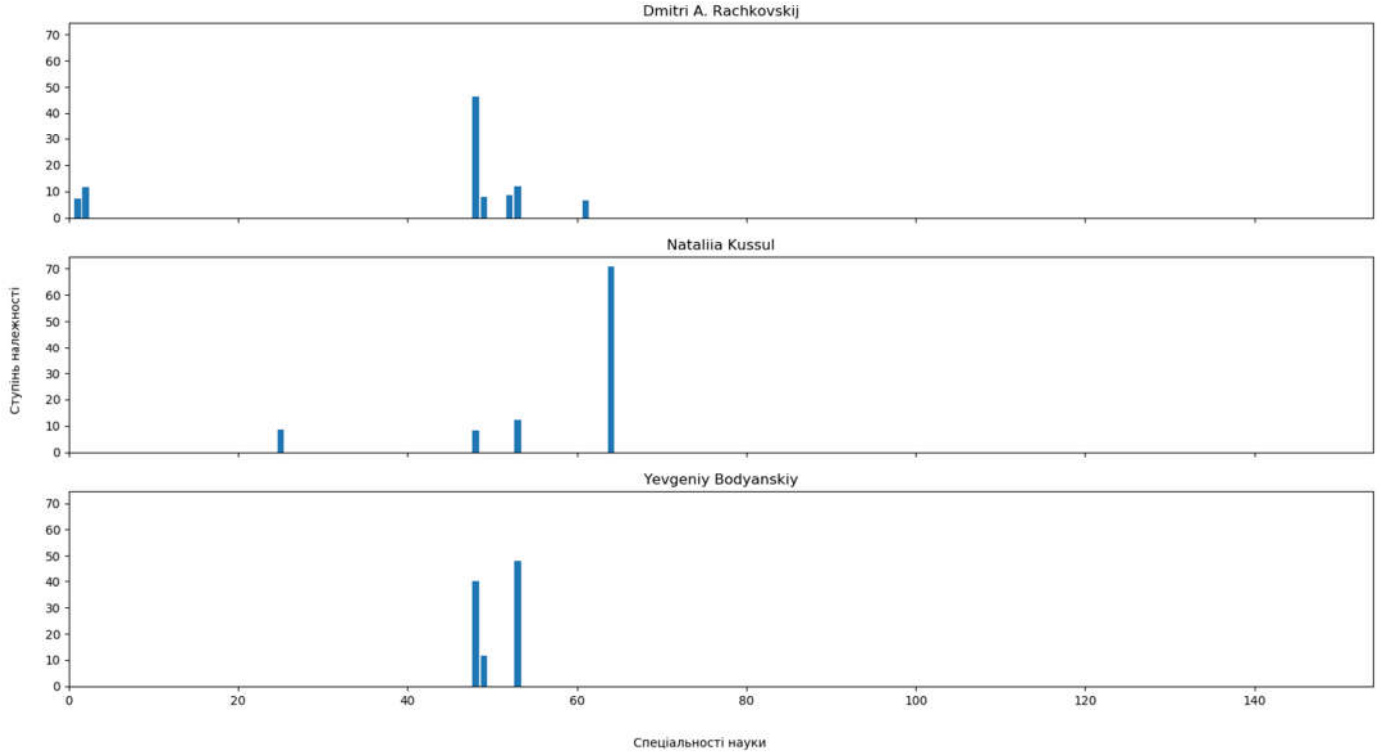


Рисунок 2.10 – Розподіл інтересів науковця після нехтування непопулярних тематик

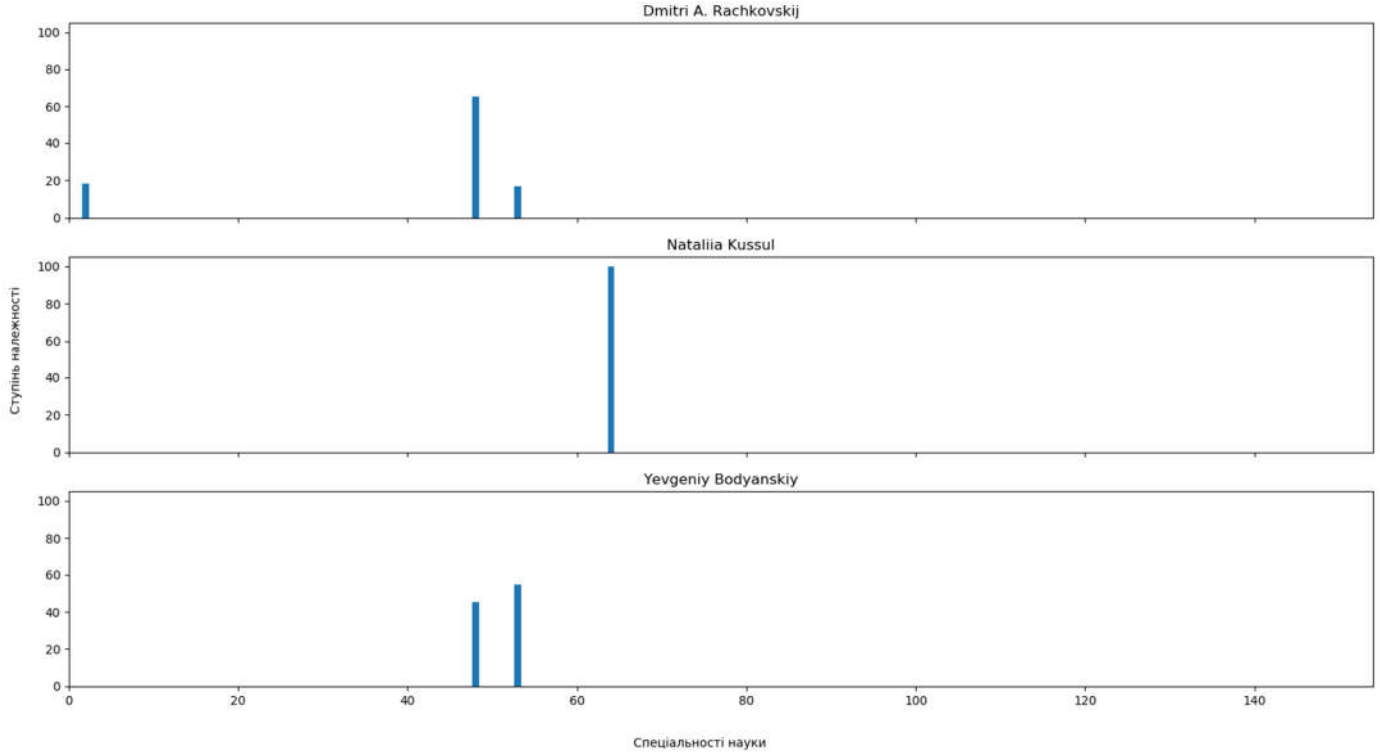


Рисунок 2.11 – Після фільтру тематик з перерахуванням ступеня належності на основі схожості спеціальностей науки

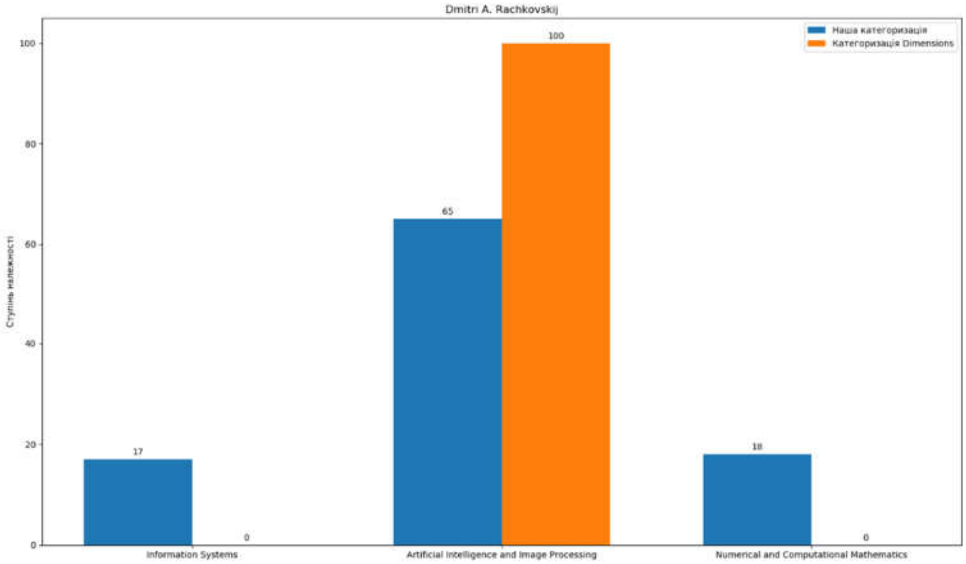


Рисунок 2.12 – Порівняння категоризації тестового науковця з категоризацією Dimensions

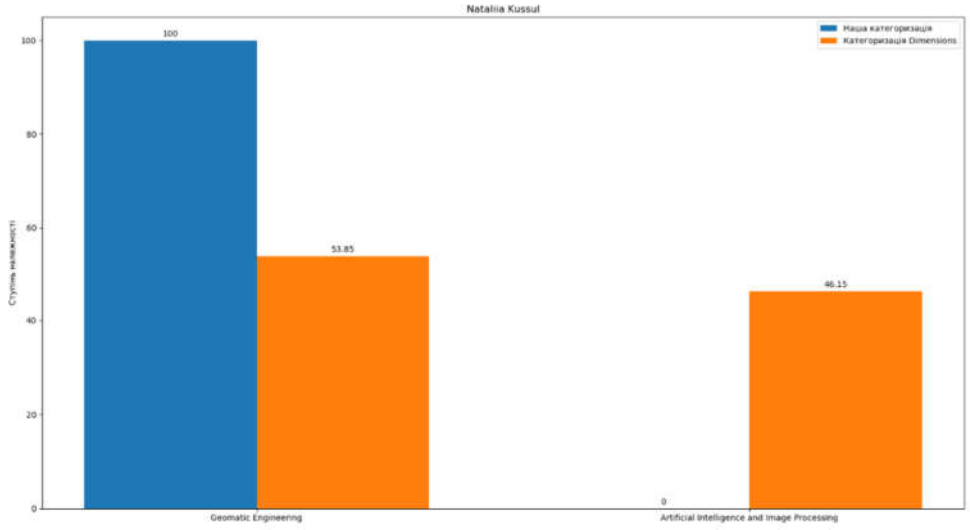


Рисунок 2.13 – Порівняння категоризації тестового науковця з категоризацією Dimensions

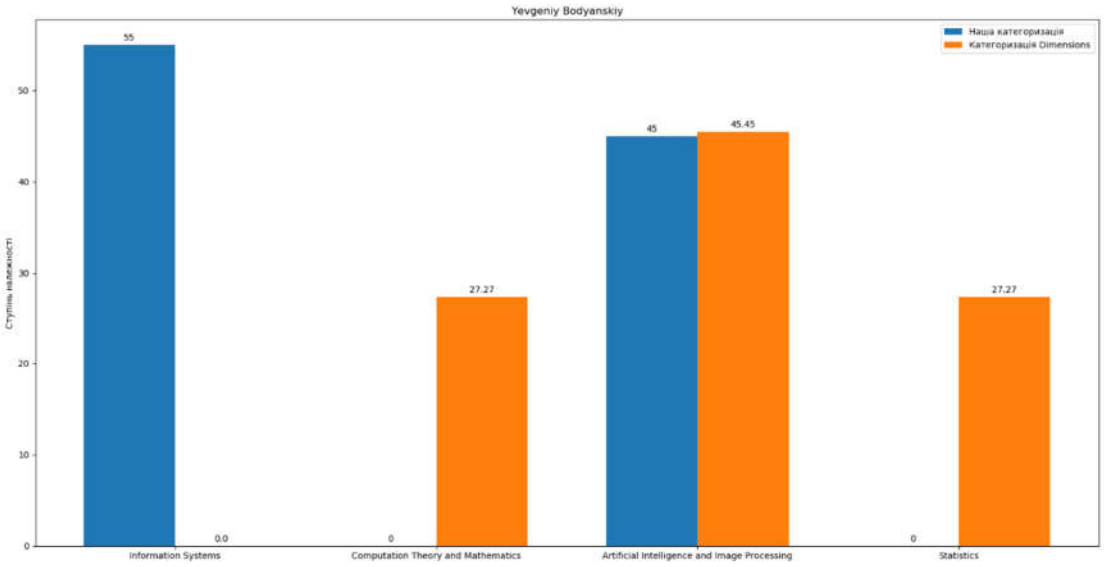


Рисунок 2.14 – Порівняння категоризації тестового науковця з категоризацією Dimensions

2.6 Визначення схожості між категоризованими науковцями

У даному підрозділі подано одне із можливих застосувань категоризації, а саме пошук «найближчих» колег. Тобто науковців, які мають схожі інтереси і за результатами категоризації віднесенні до однакових/споріднених галузей/підгалузей. Для досягнення цього необхідна метрика, міра схожості двох науковців. Маючи лише інтереси науковця (первинна, неструктурована інформація) важко сказати, наскільки два науковці є схожими. Це є ще однією причиною для необхідності категоризації науковця – визначення схожості між двома науковцями.

Як результат категоризації у нас є два вектори – вектор належності до галузей та підгалузей. Тому задача пошуку «найближчих» колег полягає у визначенні найменшої відстані між векторами належності науковців. Необхідно обрати метрику для визначення відстані між векторами. Існують різноманітні метрики. Найбільш популярною вважається косинусоїдна метрика, яка ґрунтується на базовому понятті про скалярний добуток векторів.

$$similarity = \frac{\sum_{i=1}^{n} A_i B_i}{\|A\|[\![B]\!]}, \tag{2.6}$$

де A та B – вектори.

Однією з переваг цієї метрики для нашої системи є те, що вона добре працює на «рідких» векторах, в яких більшість елементів дорівнює 0. Але оскільки ми оперуємо нечіткими множинами, дана метрика не є наочною. Для наочності у прикладах ми будемо використовувати інші метрики, хоча експерименти показали, що відстань за різними метриками майже однакова. Перша з метрик, яку ми будемо використовувати це метрика Чекановського.

$$similarity(A, B) = \frac{2\sum_{i=1}^{n} \min(A_i B_i)}{\sum_{i=1}^{n} A_i + \sum_{i=1}^{n} B_i}, \tag{2.7}$$

де A та B – вектори.

Для випадку пошуку найближчих науковців сума ступенів належності до науки дорівнює 1, тому метрику Чекановського можна записати наступним чином:

$$similarity(A,B) = \sum_{i=1}^{n} \min(A_i B_i). \quad (2.8)$$

Основний недолік метрики Чекановського – не врахування розкиду між галуззям, тобто, як сильно відрізняються відповідні значення для двох векторів. Це враховано у наступній метриці яка носить назву коефіцієнт Жаккара. Формально, коефіцієнт Жаккара для двох нечітких векторів визначається як відношення їх перетину до їх об'єднання:

$$similarity(A,B) = \frac{\sum_{i=1}^{n} \min(A_i B_i)}{\sum_{i=1}^{n} max(A_i B_i)}. \quad (2.9)$$

Обравши відповідну метрику можна визначити відстань між будь-якими науковцями. На перший погляд достатньо визначити відстань між вектором галузей але як показала практика це не завжди так. Справа в тому, що одна галузь співвідноситься з певною кількістю підгалузей. Враховуючи це науковці можуть мати спільні галузі але різні підгалузі. Тому виникає потреба в комбінованому показникові, що дозволить врахувати одразу галузі та підгалузі. Ми пропонуємо наступну метрику для визначення відстані між векторами галузей та підгалузей:

$$distance(A,B) = 1 - \big(p * GF(A,B) + (1-p) * GF(A,B)\big), \quad (2.10)$$

де GF – метрика подібності між векторами галузей, GS - метрика подібності між векторами підгалузей, $0 \leq p \leq 1$, – деякий коефіцієнт, що надає перевагу галуззям або підгалузям; метрики визначаються наступним чином:

$$GF(A,B) = similarity(A,B),$$
$$GS(A,B) = similarity(A,B),$$

Насправді, запропонований показник відстані можна дещо вдосконалити. У нашому розпорядженні також є інтереси науковцями, власне кажучи, завдяки яким нам вдалось отримати вектори галузей та підгалузей. Очевидно, що якщо науковці мають однакові інтереси (лексикографічно), то їх схожість зростає. Враховуючи це, запропонований вище показник можна записати як:

$$distance(A,B) = 1 - (k1 * GF(A,B) + k2 * GS(A,B) + GI(A,B)), \quad (2.11)$$

де $0 \leq k1, k2, k3 \leq 1$ – ваги важливості кожного показнику схожості науковців, GI – схожість науковців за інтересами.

Для прикладу роботи покажемо найближчих науковців до науковця з прикладу у описі алгоритму категоризації – Штовби С.Д. Як міру відстані оберемо відстань 4.6, встановимо p рівним 0.25 та візьмемо метрику Чекановського. Результати подано у таблиці 2.3.

Таблиця 2.3 – Список українських науковців, найближчих за інтересами до Штовби С.Д.

ID	Початковий набір ключових слів науковця	Належність до підгалузей		Відстань
		Artificial Intelligence and Image Processing	Information Systems	
3734	programming;intelligent systems	0.507	0.493	0.0011
5804	computer science; computer graphics; visualization; multimedia technology; e-learning	0.516	0.484	0.0065
12885	fuzzy sets	0.521	0.48	0.0087
963	information technology; evolutionary algorithms; natural computing	0.486	0.514	0.017
944	intelligent data analysis; mathematical modeling	0.483	0.517	0.0196

2.7 Висновки до розділу 2

Подано формалізовану постановку задачі та описано загальну структуру методу категоризації науковців у вигляді псевдокоду. Вказано основні принципи на яких базується метод.

Сформовано вимоги до сервісів баз публікацій та обгрунтовано вибір класифікації наук та бази публікацій. Протестовано метод на різних базах публікацій. Розглянуто категоризацію науковців у системі класифікації наук ANZSRC.

Проведено дослідження схожості наукових спеціальностей у системі класифікації наук ANZSRC та на основі даних з Dimensions. Схожість наукових спеціальностей використано для визначення схожості між науковцями там для категоризації науковців при відкиданні непопулярних спеціальностей. Проаналізовано можливі метрики для порівняння категоризованих науковців.

Роботу методу подано на прикладі категоризації трьох українських науковців з порівнянням з категоризацією Dimensions на основі статей.

Матеріали розділу опубліковано у працях автора [22, 47, 49].

3 ПРОГРАМНА РЕАЛІЗАЦІЯ КАТЕГОРИЗАЦІЇ

У якості перевірки роботи методу багатоярликової класифікації було розроблено програмне забезпечення класифікації українських науковців на основі системи класифікації наук Dimensions з сайту української бібліометрики за посиланням: http://nbuviap.gov.ua/bpnu/index.php. Програмне забезпечення написано на мові програмування PHP. Для збереження даних про науковців було використано систему управління базами даних MySql. Також програмне забезпечення було розроблено для Скопуса та Шпрингера.

Для роботи зі Dimensions розроблено відповідні парсери. Модуль парсингу складається з декількох ієрархій класів. Класи парсингу описуються інтерфейсом CommonParser (рис. 3.1).

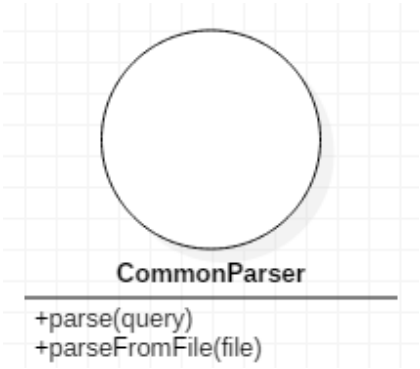


Рисунок 3.1 – Інтерфейс, що описує об’єкт «Парсер»

Інтерфейс описує операції, які може виконувати конкретний парсер. Всього два парсери. Алгоритм парсингу спільний для багатьох сайтів. Тому конкретні парсери наслідуються від абстрактного класу AbstractParser, який реалізує інтерфейс CommonParser (рис. 3.2).

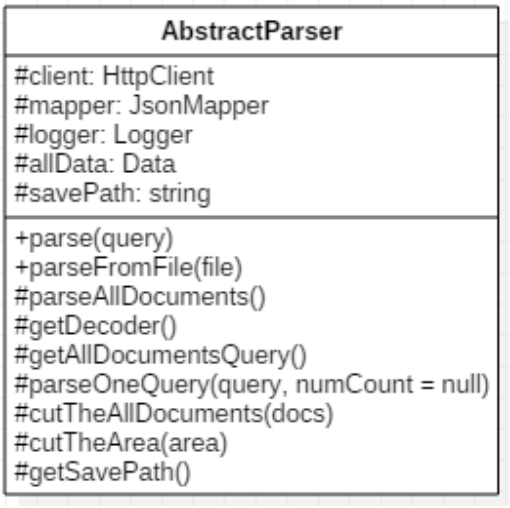


Рисунок 3.2 – Базовий парсер, що реалізує алгоритм та делегує конкретні завдання своїм нащадкам

В вказаному вище даному класі описується алгоритм парсингу. Алгоритм є спільним для деяких парсерів. Це так званий шаблон ООП (об'єктно-орієнтованого програмування) «Template method». Він реалізує операції CommonParser використовуючи абстрактні захищені методи показані вище. Будь-який парсер повинен реалізувати ці методи. Для отримання html сторінки використовується Http клієнт Guzzle який обгорнений у клас HttpClient. Для конвертування json рядку в об'єкт використовується JsonMapper. Для збереження логів використовується Logger. Реалізації CommonParser для Скопусу – зображена на рис. 3.3, а для Шпрингеру на рис. 3.4.

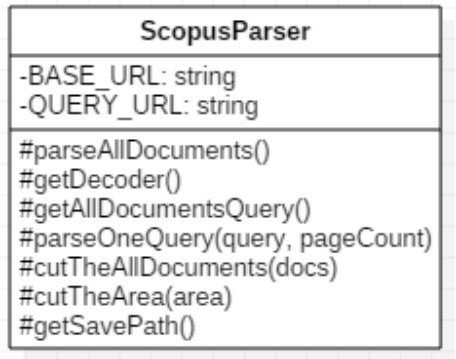


Рисунок 3.3 – Парсер сайту Scopus

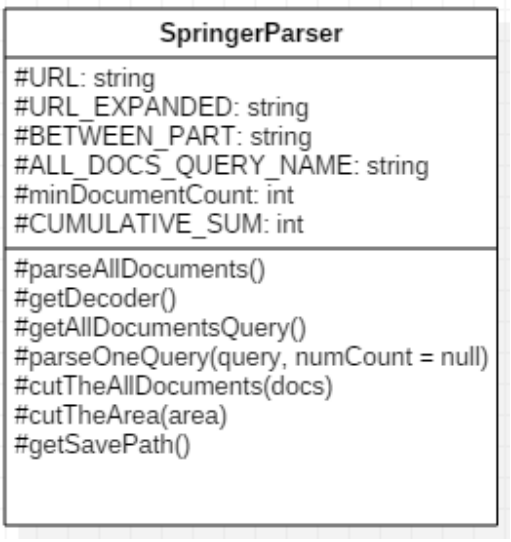


Рисунок 3.4 – Парсер сайту Springer

Інша ієрархія класів описує те, що повертається внаслідок парсингу. Ієрархію класів описує інтерфейс Data (рис.3.5).

Рисунок 3.5 – Інтерфейс, що описує базовий тип даних, що повертається після парсингу

Після виконання парсингу конкретний парсер обгортає дані у відповідну реалізацію інтерфейсу Data (рис. 3.6 та 3.7). Загальна діаграма класів модулю парсингу наведена на рис. 3.8.

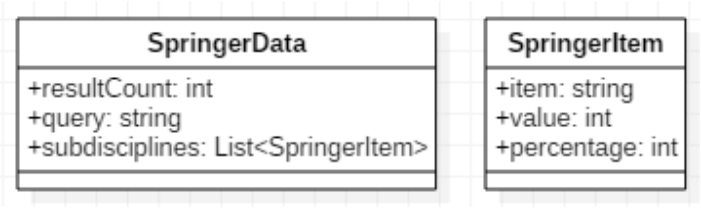


Рисунок 3.6 – Обгортка даних, що повертаються після парсингу об’єктом SpringerParser

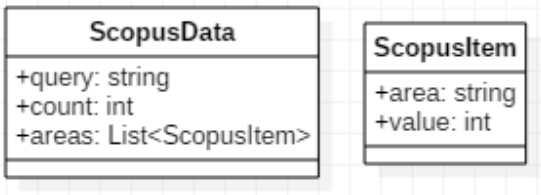


Рисунок 3.7 – Обгортка даних, що повертаються після парсингу об’єктом ScopusParser

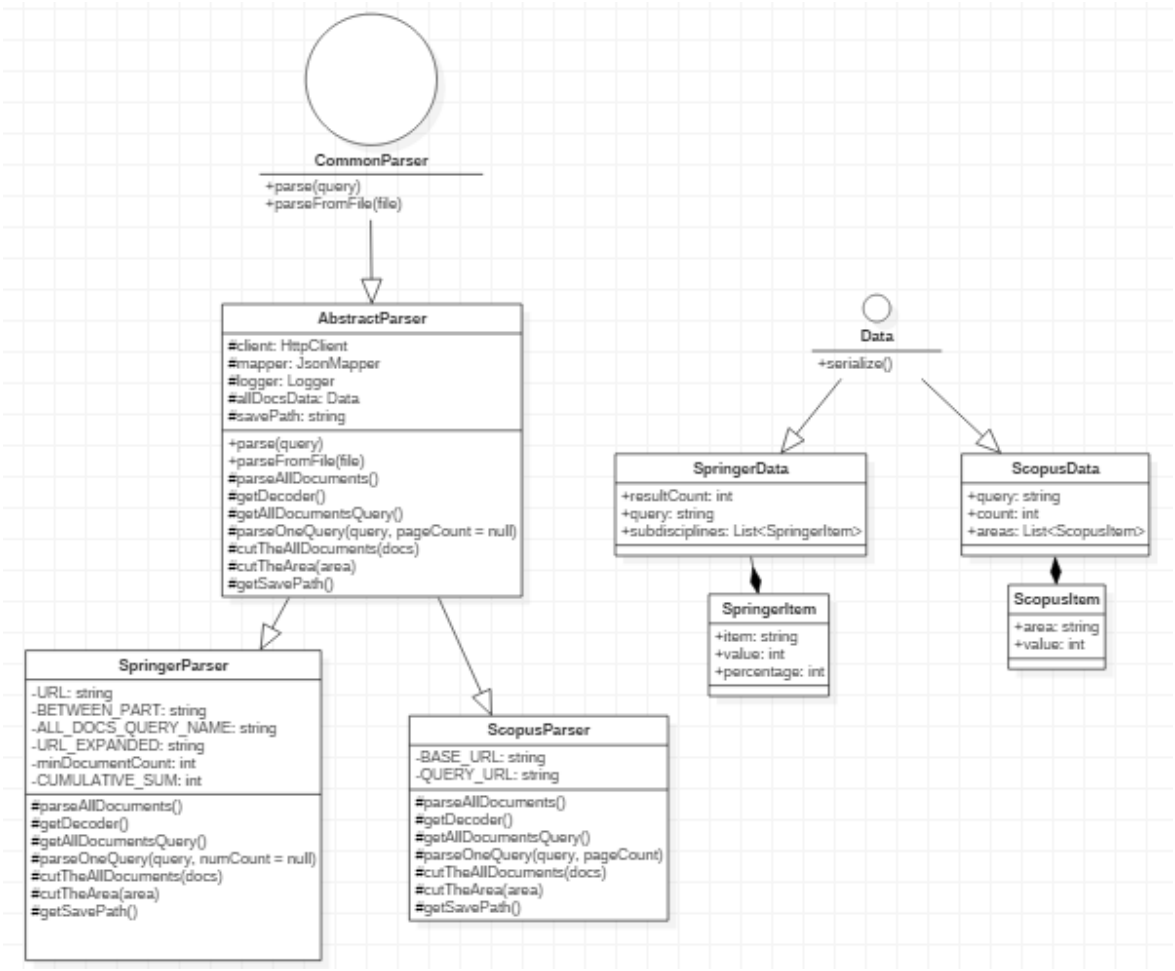


Рисунок 3.8– Діаграма класів модуля парсингу

Окрім модуля парсингу реалізовано модуль алгоритму, який опрацьовує список отриманої кількості документів за кожним інтересом науковця за алгоритмом з розділу 2. Реалізовано два конкретних алгоритми: алгоритм опрацювання результатів зі Скопусу та Шпрингера. Вони мають спільні риси тому під час їх реалізації повторювані методи були винесені у абстрактний клас. Діаграма класів

подана на рисунку 3.9. В даному модулі алгоритм описується інтерфейсом Algorithm з методом getFieldNameValue(data), що приймає параметр data з типом Data з попередньої діаграми. Для зручності увесь список піддисциплін Шпрингера та галузей Скопуса конвертується за допомогою методу prepareItems(items) у асоціативний масив, де ключем є назва галузі/піддисципліни а значенням – відповідна кількість документів.

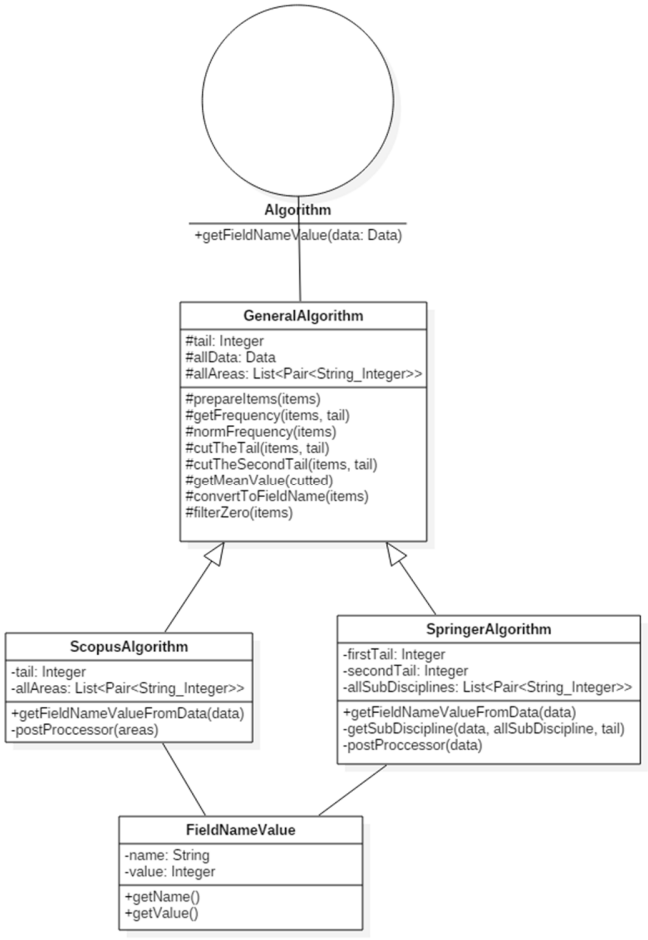


Рисунок 3.9 – Діаграма класів модулю алгоритму

В даній роботі ми працюємо з профілями науковців з Google Scholar. Структура такого науковця має вигляд поданий на рисунку 3.10.

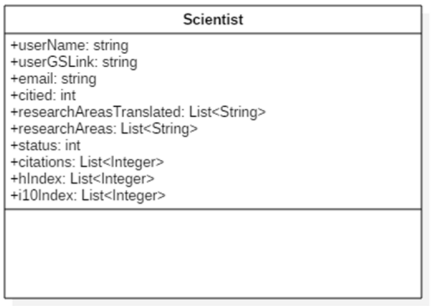


Рисунок 3.10 – Структура профіля науковця з Google Scholar

Модуль алгоритму працює лише з статистичними даними, тому для отримання галузей/піддисциплін по науковцю розроблено модуль класифікації вчених за науковими інтересами, що використовує модуль алгоритму описаний вище. Діаграма класів даного модулю має вигляд поданий на рисунку 3.11. Даний модуль не оновлює науковців з їх галузями/піддисциплінами в базі даних а лише повертає список галузей/піддисциплін характерних для даного науковця для подальшої обробки.

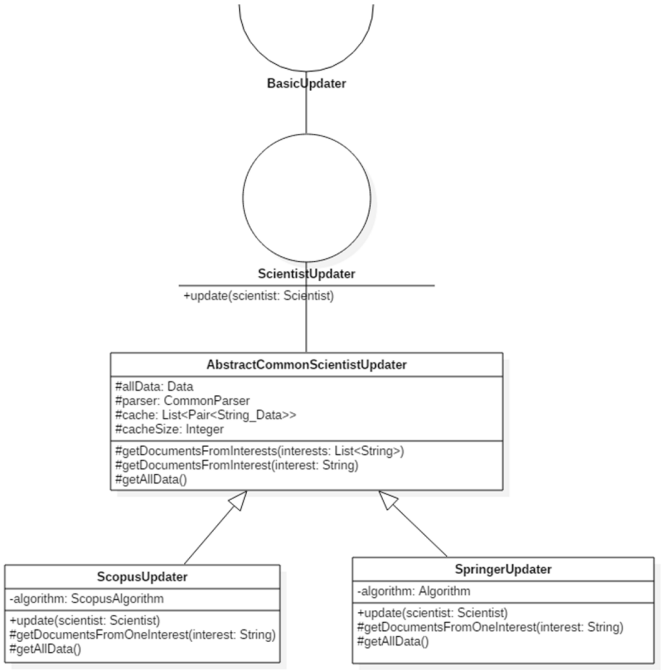


Рисунок 3.11 – Діаграма класів модулю категоризації вчених

Окрім цих модулів було розроблено Web інтерфейс для роботи з системою. Даний інтерфейс надає інформацію про усіх категоризованних науковців та дозволяє сортувати їх за різними критеріями. Інтерфейс поки що не дає прямого доступу до API модулів категоризації описаних вище. Web інтерфейс побудований на технології jQuery з використанням ajax для відправки асинхронних запитів до серверу. Сервер на деякі запити генерує html сторінки а на деякі повертає json у відповідь з інформацією про науковців. Для оброки запитів на сервері було розроблено модуль маршрутизації отриманих запитів. За своєю суттю він нагадує шаблон проектування «Ланцюжок обов'язків». Діаграма класів цього модулю подана на рисунку 3.12. Модуль складається з декількох сутностей: Action (обробник), Router (маршрутизатор), ActionFactory (фабрика обробників). Будь-який запит потрапляє до маршрутизатора FirstRouter, який містить у собі (через батька) фабрику RouterFactory з обробниками на запит з ключем маршрутизації / та інші маршрутизатори, наприклад, AlgorithmRouter. Під ключем маршрутизації можна розуміти шлях до ресурсу в домені. Наприклад, у домені http://localhost/profiles.com/algorithm ключем маршрутизації є algorithm, відповідно маршрутизатор, який має такий ключ може перенаправляти будь-який наступний запит: http://localhost/profiles.com/algorithm/*. Усю логіку запиту виконують обробники. На діаграмі показано лише три обробники, які виконують запити домашньої сторінки (HomeAction), оновлення галузей скопусу (UpdateScientisScopusAction) та шпрингера (UpdateScientistSpringerAction) для науковців.

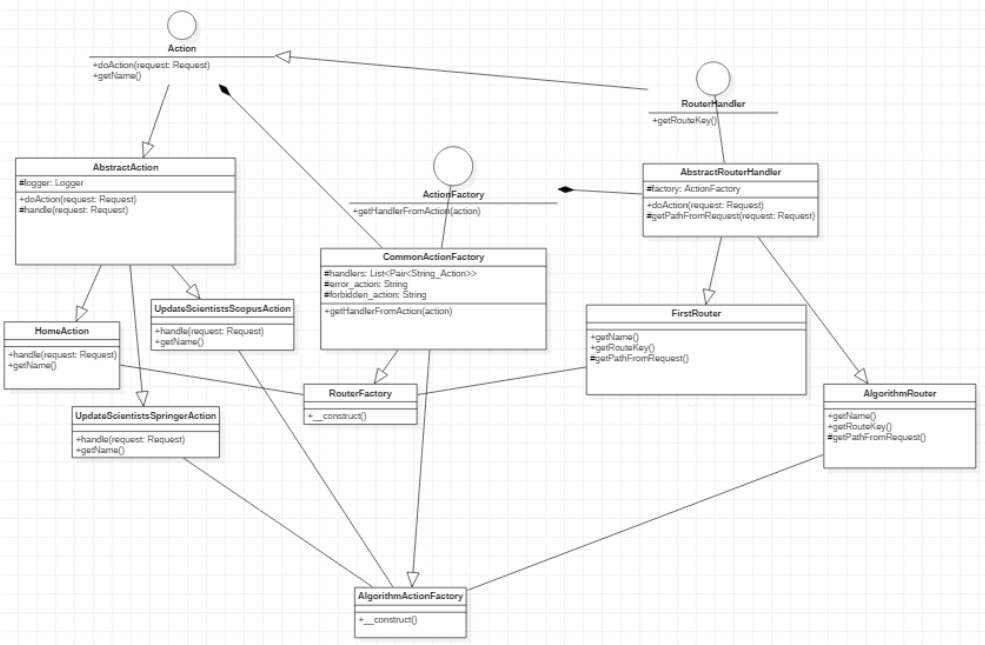


Рисунок 3.12 – Частина модулю маршрутизації запитів

4 ЕКОНОМІЧНА ЧАСТИНА

4.1 Оцінювання комерційного потенціалу розробки

Метою проведення технологічного аудиту є оцінювання комерційного потенціалу розробки інформаційної системи ієрархічної категоризації науковців за відкритими даними з Google Scholar, яка підвище точність та продуктивність існуючого підходу до категоризації науковців.

Для проведення технологічного аудиту було залучено 3-х незалежних експертів Вінницького національного технічного університету, кафедри комп'ютерних систем управління. За допомогою таблиці 4.1 за п'ятибальною шкалою використовуючи 12 критеріїв оцінки комерційного потенціалу розробки експерти надали свої оцінки.

Таблиця 4.1 – Рекомендовані критерії оцінювання комерційного потенціалу розробки та їх можлива бальна оцінка

Критерії оцінювання та бали (за 5-ти бальною шкалою)					
Кри-терій	0	1	2	3	4
Технічна здійсненність концепції:					
1	Достовірність концепції не підтверджена	Концепція підтверджена експертними висновками	Концепція підтверджена розрахунками	Концепція перевірена на практиці	Перевірено роботоздатність продукту в реальних умовах
Ринкові переваги (недоліки):					
2	Багато аналогів на малому ринку	Мало аналогів на малому ринку	Кілька аналогів на великому ринку	Один аналог на великому ринку	Продукт не має аналогів на великому ринку

Продовження табл. 4.1

3	Ціна продукту значно вища за ціни аналогів	Ціна продукту дещо вища за ціни аналогів	Ціна продукту приблизно дорівнює цінам аналогів	Ціна продукту дещо нижче за ціни аналогів	Ціна продукту значно нижче за ціни аналогів
4	Технічні та споживчі властивості продукту значно гірші, ніж в аналогів	Технічні та споживчі властивості продукту трохи гірші, ніж в аналогів	Технічні та споживчі властивості продукту на рівні аналогів	Технічні та споживчі властивості продукту трохи кращі, ніж в аналогів	Технічні та споживчі властивості продукту значно кращі, ніж в аналогів
5	Експлуатаційні витрати значно вищі, ніж в аналогів	Експлуатаційні витрати дещо вищі, ніж в аналогів	Експлуатаційні витрати на рівні експлуатаційних витрат аналогів	Експлуатаційні витрати трохи нижчі, ніж в аналогів	Експлуатаційні витрати значно нижчі, ніж в аналогів
Ринкові перспективи					
6	Ринок малий і не має позитивної динаміки	Ринок малий, але має позитивну динаміку	Середній ринок з позитивною динамікою	Великий стабільний ринок	Великий ринок з позитивною динамікою
7	Активна конкуренція великих компаній на ринку	Активна конкуренція	Помірна конкуренція	Незначна конкуренція	Конкурентів немає
Практична здійсненність					
8	Відсутні фахівці як з технічної, так і з комерційної реалізації ідеї	Необхідно наймати фахівців або витрачати значні кошти та час на навчання наявних фахівців	Необхідне незначне навчання фахівців та збільшення їх штату	Необхідне незначне навчання фахівців	Є фахівці з питань як з технічної, так і з комерційної реалізації ідеї

Продовження табл. 4.1

9	Потрібні значні фінансові ресурси, які відсутні. Джерела фінансування ідеї відсутні	Потрібні незначні фінансові ресурси. Джерела фінансування відсутні	Потрібні значні фінансові ресурси. Джерела фінансування є	Потрібні незначні фінансові ресурси. Джерела фінансування є	Не потребує додаткового фінансування
10	Необхідна розробка нових матеріалів	Потрібні матеріали, що використовуються у військово-промисловому комплексі	Потрібні дорогі матеріали	Потрібні досяжні та дешеві матеріали	Всі матеріали для реалізації ідеї відомі та давно використовуються у виробництві
11	Термін реалізації ідеї більший за 10 років	Термін реалізації ідеї більший за 5 років. Термін окупності інвестицій більше 10-ти років	Термін реалізації ідеї від 3-х до 5-ти років. Термін окупності інвестицій більше 5-ти років	Термін реалізації ідеї менше 3-х років. Термін окупності інвестицій від 3-х до 5-ти років	Термін реалізації ідеї менше 3-х років. Термін окупності інвестицій менше 3-х років
12	Необхідна розробка регламентних документів та отримання великої кількості дозвільних документів на виробництво та реалізацію продукту	Необхідно отримання великої кількості дозвільних документів на виробництво та реалізацію продукту, що вимагає значних коштів та часу	Процедура отримання дозвільних документів для виробництва та реалізації продукту вимагає незначних коштів та часу	Необхідно тільки пові-домлення відповідним органам про виробництво та реалізацію продукту	Відсутні будь-які регламентні обмеження на виробництво та реалізацію продукту

Таблиця 4.2 – Рівні комерційного потенціалу розробки

Середньоарифметична сума балів СБ, розрахована на основі висновків експертів	Рівень комерційного потенціалу розробки
0-10	Низький
11-20	Нижче середнього
21-30	Середній
31-40	Вище середнього
41-48	Високий

В таблиці 4.3 наведено результати оцінювання експертами комерційного потенціалу розробки.

Таблиця 4.3 – Результати оцінювання комерційного потенціалу розробки

Критерії	Прізвище, ініціали, посада експерта		
	Експерт 1	Експерт 2	Експерт 3.
	Бали, виставлені експертами:		
1	3	2	2
2	3	3	3
3	3	3	2
4	4	3	3
5	4	4	4
6	2	1	1
7	3	3	4
8	3	4	4
9	2	2	3
10	4	4	4
11	4	4	4
12	4	4	4
Сума балів	$СБ_1$=39	$СБ_2$=37	$СБ_3$=38
Середньоарифметична сума балів $\overline{СБ}$	$\overline{СБ} = \frac{\sum_{1}^{3} СБ_i}{3} = \frac{39+37+38}{3} = 38$		

Середньоарифметична сума балів, розрахована на основі висновків експертів склала 38, що згідно таблиці 4.2 вважається, що рівень комерційного потенціалу розробки є вище середнього.

Розробка буде реалізована в університеті та буде виставлена на його сайті. Ймовірні користувачі: науковці, студенти для пошуку інших науковців зі схожими науковими інтересами, для пошуку рецензентів робіт, для пошуку наукових керівників, для пошуку фахівців певної галузі.

Проведемо порівняння розробки, яка розробляється з аналогом, який існує на ринку. В якості аналога для розробки було обрано бібліометрику української науки. Основними недоліками аналога є суб'єктивність при здійсненні категоризації. Також до недоліків можна віднести обмеженість у системі наук та однозначна відповідність між науковцем та галуззю науки

У розробці дана проблема вирішується шляхом застосування методів штучного інтелекту. Також система випереджає аналог за такими параметрами як – швидкодія, кількість людських ресурсів.

В таблиці 4.4 наведені основні техніко-економічні показники аналога і нової розробки.

Таблиця 4.4 - Основні технічні показники аналога і нової розробки

Показники	Аналог	Нова розробка	Відношення параметрів нової розробки до параметрів аналога
Кількість галузей науки	12	22	1.83
Швидкодія	1	10	10
Витрати на виконання роботи	1	10	10

Продовження табл. 4.4

Кількість підгалузей	180	154	0.85
Нерівномірність розподілу	1	2	2

Виходячи із даних таблиці 4.4 можна зробити висновок, що у зв'язку з швидким розвитком науки і техніки все більше і більше набуває популярності автоматизація різноманітних процесів. Саме цей показник відсутній у системі конкурента – усі дії пов'язані з категоризацією виконуються людиною, тобто суб'єктивно, з певними затратами часу та більш затратно у порівнянні з автоматичною системою.

4.2 Прогнозування витрат на виконання науково-дослідної роботи

Проведемо прогнозування витрат на виконання робіт у три етапи:

*1-й етап:*Розрахунок витрат, які безпосередньо стосуються виконавців даного розділу НДР.

2-й етап: Розрахунок загальних витрат на виконання НДР;

3-й етап: Прогнозування загальних витрат на виконання та впровадження НДР.

Проведемо перший етап прогнозування.

1. Основна заробітна плата кожного із дослідників $З_О$, якщо вони працюють в наукових установах бюджетної сфери визначається за формулою:

$$З_0 = \frac{М}{Т_р} * t \text{ (грн)} \quad (4.1)$$

де *М* – місячний посадовий оклад конкретного розробника (інженера, дослідника, науковця тощо), грн.;

T_P – число робочих днів в місяці; приблизно $T_P \approx 21...23$ дні;

t – число робочих днів роботи дослідника.

Дану розробку буде проводити інженер-програміст, величина окладу буде становити 6500 грн. на місяць. Кількість робочих днів у місяці складає 22, а кількість робочих днів дослідника складає 55. Зведемо сумарні розрахунки до таблиця 4.5.

Таблиця 4.5 – Заробітна плата дослідника в науковій установі бюджетної сфери

Найменування посади	Місячний посадовий оклад, грн.	Оплата за робочий день, грн.	Число днів роботи	Витрати на заробітну плату грн.
Керівник проекту	12000	545.5	5	2727
Інженер-програміст	6500	295.5	55	16250
Всього				18977

2. Розрахунок додаткової заробітної плати робітників

Додаткова заробітна плата $З_д$ всіх розробників та робітників, які приймали устають в розробці нового технічного рішення розраховується як 10 - 12 % від основної заробітної плати робітників. На даному підприємстві додаткова заробітна плата начисляється в розмірі 10% від основної заробітної плати.

$$З_д = 0{,}11 * 18977 = 2087{,}5 \text{ (грн)} \quad (4.2)$$

3. Нарахування на заробітну плату $Н_{ЗП}$ дослідників та робітників, які брали участь у виконанні даного етапу роботи, розраховуються за формулою (4.3):

$$Н_{ЗП} = (З_о + З_д) * \frac{\beta}{100} \text{ (грн)} \quad (4.3)$$

де $З_о$ – основна заробітна плата розробників, грн.;

$З_д$ – додаткова заробітна плата всіх розробників та робітників, грн.;

β – ставка єдиного внеску на загальнообов'язкове державне соціальне страхування, % .

Дана діяльність відноситься до бюджетної сфери, тому ставка єдиного внеску на загальнообов'язкове державне соціальне страхування буде складати 22%, тоді**:**

$$Н_{ЗП} = (18977 + 2087{,}5) * \frac{22}{100} = 4634{,}3 \text{ (грн)}$$

4. Амортизація обладнання, комп'ютерів та приміщень, які використовувались під час виконання даного етапу роботи

Дані відрахування розраховують по кожному виду обладнання, приміщенням тощо.

$$A = \frac{Ц \cdot T}{Tкор \cdot 12} \quad [грн], \tag{4.4}$$

де Ц – балансова вартість даного виду обладнання (приміщень), грн.;

$Т_{кор}$ – час користування;

Т – термін використання обладнання (приміщень), цілі місяці.

Згідно пункта 137.3.3 Податкового кодекса амортизація нараховується на основні засоби вартістю понад 2500 грн. В нашому випадку для написання магістерської роботи використовувався персональний комп'ютер артістю 25000 грн.

$$А = \frac{25000 \cdot 2}{2 \cdot 12} = 2083{,}3$$

5. Норма витрат матеріалу – це плановий показник, який визначає максимально допустимі затрати відповідних ресурсів на виробництво одиниці продукції в умовах певного рівня техніки і організації виробництва.

Витрати на матеріали М, що були використані під час виконання даного етапу роботи, розраховуються по кожному виду матеріалів за формулою:

$$М = \sum_{1}^{n} Н_i \cdot Ц_i \cdot К_i - \sum_{1}^{n} В_i \cdot Ц_в \quad \text{грн.}, \tag{4.5}$$

де $Н_i$ – витрати матеріалу *і*-го найменування, кг;

$Ц_і$ – вартість матеріалу *і*-го найменування, грн./кг.;

$К_і$ – коефіцієнт транспортних витрат, $К_і$ = (1,1…1,15);

$В_і$ – маса відходів матеріалу *і*-го найменування, кг;

$Ц_в$ – ціна відходів матеріалу *і*-го найменування, грн/кг;

n – кількість видів матеріалів.

Інформацію про використані матеріали подамо у вигляді табл. 4.6.

Таблиця 4.6– Матеріали, що використані на розробку

Найменування матеріалу	Ціна за одиницю, грн.	Витрачено	Вартість витраченого матеріалу, грн.
Папір	75	1	75
Ручка	12	1	12
Флешка	120	1	120
CD-диск	15	1	15
З врахуванням коефіцієнта транспортування			244.2

6. До статті «Паливо та енергія на технологічні цілі» відносяться витрати на всі види палива й енергії, що безпосередньо використовуються у процесі виробництва продукції. У даному випадку будемо враховувати лише витрати на електроенергію, яку споживає комп'ютер:

$$В_е = В \cdot П \cdot Ф \cdot К_п, \quad (4.6)$$

де В – вартість 1 кВт енергії, грн. В = 8,44 грн/кВт*год;

П – установлена потужність обладнання, кВт. П = 500 Вт або П = 0,5 кВт;

Ф – фактична кількість годин роботи обладнання, год. Ф = 200 год.;

$К_п$ – коефіцієнт використання потужності, $К_п$ = 0,65.

$$В_е = 8{,}44 \cdot 0{,}5 \cdot 200 \cdot 0{,}65 = 548{,}6 (грн).$$

Інші витрати $B_{ін}$ охоплюють: витрати на управління організацією, оплата службових відряджень, витрати на утримання, ремонт та експлуатацію основних засобів, витрати на опалення, освітлення, водопостачання, охорону праці тощо. Інші витрати $B_{ін}$ можна прийняти як (100…300)% від суми основної заробітної плати розробників та робітників, які виконували дану МКНР, тобто:

$$B_{ін} = (1..3) \cdot (З + З_р). \tag{4.7}$$

$$B_{ін} = 1 \cdot 18977 = 18977\,(грн.)$$

Сума всіх попередніх статей витрат дає витрати, які безпосередньо стосуються даного розділу МКНР

$$B = 18977 + 2087,5 + 4634,3 + 2083,3 + 244,2 + 548,6 + 18977 = 47552,4\,(грн.)$$

Загальна вартість всієї МКНР визначається за формулою:

$$B_{заг} = \frac{B}{\alpha} \tag{4.8}$$

$$B_{заг} = \frac{47552,4}{1} = 47552,4\,(грн.)$$

Прогнозування загальних втрат ЗВ на виконання та впровадження результатів виконаної МКНР здійснюється за формулою:

$$ЗВ = \frac{B}{\beta}, \tag{4.9}$$

де $\beta -$ коефіцієнт, який характеризує стадію виконання даної НДР.

Оскільки, робота знаходиться на стадії науково-дослідних робіт, то коефіцієнт $\beta = 0,9$.

Звідси:

$$ЗВ = \frac{47552,4}{0,9} = 52836(грн.).$$

4.3 Оцінка внеску НДР

Для цього спочатку потрібно визначити коефіцієнт наукової значимості отриманих результатів НДР, а потім підрахувати внесок в досягнення цих результатів.

Коефіцієнт наукової значимості результатів проведеної науково-дослідної роботи $К_{ЗН}$ можна підрахувати за формулою:

$$К_{ЗН} = \frac{\sum_1^3 b_i * d_i}{\sum_1^3 b_{max} * d_i} \qquad (4.10)$$

де b_i – значимість отриманих результатів: b_1 – ступінь наукової новизни, b_2 – рівень теоретичної обґрунтованості, b_3 – ступінь експериментальної перевірки результатів. Бальна оцінка отриманих результатів наведена в таблиці 4.8. Максимальне значення отриманих за кожною характеристикою результатів можна прийняти в межах 7...10 балів;

d_i – питома вага кожної характеристики, значення якої наведені в таблиці 4.8;

3 – кількість характеристик, за якими була зроблена оцінка результатів науково-дослідної роботи.

Таблиця 4.8 – Показники для оцінювання наукової значимості результатів виконаної НДР

Характеристики	Питома вага характеристики d_i	Бальна оцінка характеристики		
		Ступінь новизни b_1	Рівень теоретичної обґрунтованості b_2	Ступінь експериментальної перевірки результатів b_3
		1	2-5	7-10
b_1	0,500	Часткове удосконалення виробів, технологій, матеріалів, програмного продукту тощо	Суттєве удосконалення виробів, технологій, матеріалів, програмного продукту тощо	Нові напрямки в розробці виробів, технологій, матеріалів, програмного продукту тощо. Створення принципово нової техніки
b_2	0,333	Позитивне рішення на основі зроблених узагальнень	Установлення залежносте, які використовувались в інших випадках	Відкриття нових шляхів рішення задачі
b_3	0,167	Експериментальна перевірка не робилась	Результати перевірялись на невеликій кількості даних	Результати перевірені на великій кількості даних

В ході виконання даної магістерської кваліфікаційної роботи було розроблено інформаційну систему ієрархічної категоризації науковців за відкритими даними з Google Scholar, що підвищить точність та продуктивність існуючого підходу до категоризації науковців.

Тоді згідно таблиці 4.8$b_1 = 4$. Отримані результати підтвердили залежності та характеристики, отримані раніше для окремих випадків рішення даної задачі, тому$b_2 = 7$. Оскільки результати перевірялись на отриманих попередньо

експериментальних даних, приймаємо$b_3 = 5$. Тоді коефіцієнт наукової значимості результатів буде наступним

$$K_{ЗН} = \frac{4 \cdot 0,5 + 7 \cdot 0,333 + 5 \cdot 0,167}{10 \cdot 0,5 + 10 \cdot 0,333 + 10 \cdot 0,167} = 0,52 .$$

Зважаючи на малий час, відведений на проведення дослідження, отримане значення коефіцієнта наукової значимості можна вважати цілком задовільним.

Внесок дипломника в досягнення отриманих результатів НДР можна розрахувати за формулою:

$$V = \frac{k_{\text{ТВІ}} * З_i}{\sum_i^n k_{\text{ТВІ}} * З_i} \qquad (4.11)$$

де $k_{\text{ТВІ}}$ – коефіцієнт творчої участі кожного виконавця НДР, який оцінюється таким чином: проведення досліджень – 3 бали, робоче проектування – 1,5 бали, освоєння – 1,0 балів. якщо виконавець приймав участь в декількох видах робіт, то береться сума відповідних балів;

$З_i$ – заробітна плата кожного виконавця НДР, грн.;

n – кількість всіх виконавців науково-дослідної роботи: наукові керівники, відповідальні виконавці, спеціалісти, наукові співробітники, робітники різних професій тощо.

$$V = \frac{3 * 12000}{3 * 12000 + 6500 * 1,5} = 0,77$$

4.4 Висновки до розділу 4

В даному розділі було оцінено комерційний потенціал розробки інформаційної системи ієрархічної категоризації науковців за відкритими даними з Google Scholar, яка підвищить точність та продуктивність існуючого підходу до категоризації науковців. Комерційний потенціал розробки виявився на вище середньому рівні.

Прогнозування витрат на виконання науково-дослідної роботи по кожній з статей витрат складе 47552,4 грн. Загальна ж величина витрат на виконання та впровадження результатів даної НДР буде складати 52836 грн.

Отримане значення коефіцієнта наукової значимості ($K_{ЗН} = 0,52$) можна вважати цілком задовільним, оскільки час відведений на проведення дослідження незначний. Особистий внесок дипломника в наукове дослідження становить 0,77, що являється нормальним значенням.

ВИСНОВКИ

Мета роботи досягнута – створено інформаційну систему категоризації науковців за спеціальностями на основі їх наукових інтересів в Google Scholar. Категоризація здійснюється за австралійсько-новозеландською схемою класифікації наук ANZSRC з використанням інформаціних ресурсів системи Dimensions. Під час виконання роботи вирішено такі задачі:

- розроблено модель представлення науковців у просторі галузей та спеціальностей наук. Науковець співставляється із декількома спеціальностями наук (із 154 можливих) та через спеціальності визначаються галузі науковця (22 можливих);
- досліджено 4 системи класифікації наук та обгрунтувано вибір ANZSRC. для вирішення задачі категоризації;
- розроблено модель їх оцінювання схожості наукових спеціальностей на основі індекса Жакара та експериментально розраховано схожість для 10878 пар спеціальностей. Вказано можливе застосування схожості спеціальностей а саме категоризація науковця (при редукції спеціальності) та визначення схожості між науковцями (врахування схожості між спеціальностями науковців);
- покращено алгоритм категоризації науковців за спеціальностями, який враховує взаємодію як інтересів науковців, так і наукових спеціальностей. Множина інтересів науковця доповнюється сполученням інтересів між собою. Нехтування незначних спеціальностей поширює свій внесок на споріднені спеціальностей;
- реалізовано програмно запропоновані моделі та алгоритми.

Можливе застосування категоризації науковців є пошук партнерів за спільними інтересами. Така задача може виникнути при відборі науковців для рецензування або опонування дисертацій а також для виконання спільних проектів. Окрім цього дана інформація може бути використана для проведення статистичних досліджень про вид діяльності науковців.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Lopes GR, Moro MM, Wives LK, De Oliveira JPM (2010) Collaboration recommendation on academic social networks. In: Advances in Conceptual Modeling–Applications and Challenges, Springer. pp. 190–199.

2. Xiangjie Kong, Huizhen Jiang, Zhuo Yang, Zhuo Yang, Zhuo Yang (2016) Amr Tolba Exploiting Publication Contents and Collaboration Networks for Collaborator Recommendation – PlosOne.

3. G. Adomavicius and A. Tuzhilin, “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,” IEEE transactions on knowledge and data engineering, vol. 17, no. 6, pp. 734–749, 2005.

4. Kuncheva L. Combining pattern classifiers: methods and algorithms / L. Kuncheva. – John Wiley & Sons, 2004. – 350 p.

5. К.В. Воронцов Модификации EM-алгоритма для вероятностного тематического моделирования / К.В. Воронцов, А.А. Потапенко – Машинное обучение и анализ даных, 2013 – с.657-686.

6. К.В. Воронцов Вероятностное тематическое моделирование / К.В. Воронцов – 2013.

7. John M. Using Google Scholar institutional level data to evaluate the quality of university research / M. John, R. Jesse, O. Musbaudeen. // Scientometrics. – 2017. – №113. – С. 1627–1643.

8. Alberto M. A novel method for depicting academic disciplines through Google Scholar Citations: The case of Bibliometrics / M. Alberto, O. Enrique, D. Emilio. // Scientometrics. – 2018. – №114. – С. 1251–1273.

9. Бібліометрика української науки [Електронний ресурс] – Режим доступ до ресурсу: http://www.nbuviap.gov.ua/bpnu.

10. Design Patterns: Elements of Reusable Object-Oriented Software / G.Erich, H. Richard, J. Ralph, V. John., 1994. – (Addison-Wesley).

11. MySql Documentation [Електронний ресурс] – Режим доступ до ресурсу: https://dev.mysql.com/doc/.

12. Php Documentation [Електронний ресурс] – Режим доступ до ресурсу: http://php.net/docs.php.

13. Apache HTTP Server Version 2.4 Documentation [Електронний ресурс] – Режим доступ до ресурсу: https://httpd.apache.org/docs/2.4/.

14. Using Google Scholar in research evaluation of humanities and social science programs: a comparison with Web of Science data / P.Ad, C. Rodrigo, V. Thed, W. Paul. // Research Evaluation. – 2015.

15. Emilio D. Using Google Scholar for Research Assessment. A New Data Source for Bibliometric Studies: Strengths versus Weaknesses / D. Emilio, O. Enrique, M. Alberto.– 2017.

16. The application of bibliometrics to research evaluation in the humanities and social sciences: an exploratory study using normalized Google Scholar data for the publications of a research institute / B.Lutz, T. Andreas, M. Werner, S. Hermann. // Journal of the Association for Information Science and Technology. – 2016. – №67. – С. 2778–2789.

17. The lost academic home: institutional affiliation links in Google Scholar Citations / O.Enrique, M. Juan, M. Alberto, D. Emilio. // Online Information Review. – 2017. – №41. – С. 762–781.

18. Enrique O. Google Scholar as a source for scholarly evaluation: a bibliographic review of database errors / O. Enrique, M. Alberto, D. Emilio. // Revista Española de Documentación Científica. – 2017. – №40. – С. 185–218.

19. Mike T. Dimensions: A Competitor to Scopus and the Web of Science? / Thelwall Mike. // Journal of Informetrics. – 2018. – №12. – С. 430–435.

20. Springer [Електронний ресурс] – Режим доступ до ресурсу: https://link.springer.com/.

21. Scopus [Електронний ресурс] – Режим доступ до ресурсу: https://www.scopus.com/.

22. Петричко М. В. Автоматична категоризація науковців за тематикою досліджень на основі профілей в Google Scholar [Електронний ресурс] / М. В. Петричко, С. Д. Штовба // ВНТУ. – 2018. – Режим доступу до ресурсу: https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2018/paper/view/5427/4433.

23. Sun, C., King, T. J., Henville, P., Marchant R. Hierarchical Word Mover Distance for Collaboration Recommender System / C. Sun, T. J. King, P. Henville, R. Marchant // Australasian Conference on Data Mining. Communications in Computer and Information Science, Springer - 2018. – V. 996. – P.289-302.

24. Winograd T. Procedures as a Representation for Data in a Computer Program for Understanding Natural Language / T. Winograd // Computer Science and Artificial Intelligence Lab (CSAIL) – 1971.

25. Johnson M., How the Statistical Revolution Changes (Computational) Linguistics / M. Johnson // Association for Computational Linguistics. – 2009. – P. 3-11.

26. Klein D., Manning C.D. Natural Language Grammar Induction using a Constituent-Context Model / D. Klein, C. D. Manning // CiteSeer – 2002.

27. Devlin J., Chang M., Lee K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M. Chang, K. Lee / Computation and Language. – 2019.

28. Peters M., Neumann M., Iyyer M., Gardner M., Clark K., Lee K., Zettlemoyer L. Deep contextualized word representations / M. Peters, M. Neumann, M. Iyyer, M. Gardner, K. Clark, L. Lee, L Zettlemoyer // Computation and Language. – 2018.

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

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

31. Pennington J., Socher R., Manning C. D. Glove: Global Vectors for Word Representation / J. Pennington, R. Socher, C.D. Manning // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2014.

32. Mikolov T., Sutskever P., Chen K., Corrado G., Dean J. Distributed Representations of Words and Phrases and their Compositionality / T. Mikolov, P. Sutskever, K. Chen, G. Corrado, J. Dean // Advances in neural information processing systems 26. – 2013.

33. Bengio Y., Ducharme R., Vincent P., Jauvin C. A Neural Probabilistic Language Model / Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin // Journal of Machine Learning Research. – 2003. - V. 3. – P. 1137-1155.

34. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval / G. Salton, C. Buckley // Information Processing & Management. – 1988. - V. 24. – P. 512-523.

35. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. / D.M. Blei, A.Y. Ng, M.I. Jordan // J. Mach. Learn. - 2003. V. 3. - P. 993–1022.

36. Hofmann, T. Probabilistic latent semantic indexing. / T. Hofmann // In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. - 1999. - pp. 50–57.

37. Arora, S., Ge, R., Moitra, A. Learning topic models - going beyond SVD. / S. Arora, R. Ge, A. Moitra // CoRR abs/1204.1956. - 2012.

38. Mikolov, T., Chen L., Corrado G., Dean J. Efficient estimation of word representations in vector space. / T. Mikolov, L. Chen, G. Corrado, J. Dean // ICLR Workshop. - 2013.

39. Jin, J., Geng, Q., Mou, H., Chen, C. Author–Subject–Topic model for reviewer recommendation / J. Jin, Q. Geng, H. Mou, C. Chen // Journal of Information Science. - 2018.

40. Gollapalli, S.D., Mitra, P., Giles, S.L., Similar researcher search in academic environments / S.D. Gollapalli, P. Mitra, S.L. Giles // Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries. - 2012.

41. Guerra, G., Quan, W., Li, L. SCOSY: A Biomedical Collaboration Recommendation System / G. Guerra, W. Quan, L. Li // Annual International Conference of the IEEE Engineering in Medicine and Biology Society. - 2018.

42. Sun, F., Liu, L., Jin, J. An Expert Recommendation Model for Academic Talent Evaluation / F. Sun, L. Liu, J. Jin // Intelligent Computing. - 2019.

43. Zhao, Y., Tang, J., Du, Z., EFCNN: A Restricted Convolutional Neural Network for Expert Finding / Y. Zhao, J. Tang, Z. Du // Advances in Knowledge Discovery and Data Mining. - 2019.

44. Al-Zboon, S.A., Tawalbeh, S., Aljarrah H. Resolving Conflict of Interests in Recommending Reviewers for Academic Publications Using Link Prediction Techniques / S.A. Al-Zboon, S. Tawalbeh, H. Aljarrah // International Conference on New Trends in Computing Sciences. - 2019.

45. Rosen-zvi, M., Chemudugunta, C., Griffiths, T.L. Learning author-topic models from text corpora / M. Rose-zvi, C. Chemudugunta, T.L. Griffiths // ACM Transactions on Information Systems. - 2010.

46. Shtovba S., Shtovba O., Petrychko M. Detection of Social Network Toxic Comments with Usage of Syntactic Dependencies in the Sentences / Proc. of the Second International Workshop on Computer Modeling and Intelligent Systems, Zaporizhzhia, Ukraine, April 15-19, 2019. CEUR Workshops Proceeding, Vol. 2353. – 2019. – P. 313-323.CMIS-2019.

47. Петричко М. В. Статистичний підхід до оцінювання подібності наукових спеціальностей в системі Dimensions [Електронний ресурс] / М. В. Петричко, С. Д. Штовба // ВНТУ. – 2019. – Режим доступу до ресурсу: https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2019/paper/view/7821/6411.

48. Tang, H., Wu, Y., Yao, J. T., Wang, G., Yao, Y. Y. CUPTRSS: A Web-based Research Support System / H. Tang, Y. Wu, J. T. Yao, G. Wang, Y. Y. Yao // Applications, Products and Services of Web-based Support Systems. - 2003.

49. Shtovba S., Petrychko M. Jaccard index – Based Assessing the Similarity of Research Fields in Dimensions / 1st International Workshop on Digital Content & Smart Multimedia 2019, Lviv. In press.

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

51. Shtovba S., Shtovba O., Petrychko M. Impact of the syntactic dependencies in the sentences on the quality of the identification of the toxic comments in the social networks / Scientific Works of Vinnytsia National Technical University, Vinnytsia, Ukraine, 2019.

ДОДАТКИ

Додаток А

(Обов’язковий)

ВНТУ

ТЕХНІЧНЕ ЗАВДАННЯ

на виконання магістерської кваліфікаційної роботи

«Інформаційна система ієрархічної категоризації науковців за відкритими даними з Google Scholar. Частина 1. Інтелектуальний аналіз даних»

Студент групи 2АКІТ-18м Петричко М.В.

“___” _____ 2019 р.

Керівник д.т.н., професор Штовба С. Д.

“___” _____ 2019 р.

Вінниця 2019

1. Назва та галузь застосування

1.1. Назва – інформаційна система ієрархічної категоризації науковців за відкритими даними з Google Scholar. Частина 1: інтелектуальний аналіз даних.

1.2. Галузь застосування – Наукометрія.

2. Підстава для проведення розробки.

Тема магістерської кваліфікаційної роботи затверджена наказом по ВНТУ № 254 від 02.10.2019р.

3. Мета та призначення розробки.

Метою магістерської кваліфікаційної роботи є розробка автоматизованої системи для категоризації науковців.

4. Вихідні дані для проведення розробки.

Магістерська кваліфікаційна робота виконується вперше. В ході проведення розробки повинні використовуватись такі документи:

1. Zhang M. L., A review on multi-label learning algorithms / M.L. Zhang, Z. H. Zhou // IEEE transactions on knowledge and data engineering. – 2014. – Vol. 26. – №.8. – P. 1819-1837.

2. Городецкий В. И. Методы и алгоритмы коллективного распознавания / Городецкий В.И., Серебряков С.В. // Автоматика и телемеханика. – 2008. – №11. – С. 3-40.

3. Kuncheva L. Combining pattern classifiers: methods and algorithms / L. Kuncheva. – John Wiley & Sons, 2004. – 350 p.

4. К.В. Воронцов Модификации EM-алгоритма для вероятностного тематического моделирования / К.В. Воронцов, А.А. Потапенко – Машинное обучение и анализ даных, 2013 – с.657-686.

5. К.В. Воронцов Вероятностное тематическое моделирование / К.В. Воронцов – 2013.

5. Вимоги до розробки.

5.1. Перелік головних функцій:

- категоризація науковців за галузями та підгалузями;
- не менше 22 галузей та 154 підгалузей;
- ранжування науковців за належністю до галузі та підгалузі;
- впорядковування науковців за показниками Google Scholar.

5.2. Основні технічні вимоги до розробки.

5.2.1. Вимоги до програмної платформи:

- WINDOWS 7\8\10;
- PHP 5.6;
- Apache;
- MySql.

5.2.2. Умови експлуатації системи:

- робота на стандартних ПЕОМ в приміщеннях зі стандартними умовами;

- можливість цілодобового функціонування системи;
- текст програмного забезпечення системи є цілком закритим.

6. Економічні показники

До економічних показників входять:
- термін окупності не більше 3 років;
- коефіцієнт наукової значимості 0.52;
- коефіцієнт внеску дипломника 0.77;
- загальні витрати не більше 55000 грн.;
- рівень комерційного потенціалу розробки вище середнього.

7. Стадії та етапи розробки.

7.1 Пояснювальна записка:

1	Вступ. Постановка задачі дослідження.	10.09.2019р.
2	Розробка методу категоризації..	20.10.2019р.
3	Програмна реалізація.	5.11.2019р.
4	Підготовка економічної частини	15.11.2019р.
5	Апробація результатів дослідження	22.11.2019р.
6	Публікації	1.12.2019
7	Оформлення пояснювальної записки, графічного матеріалу і презентації	10.12.2019р.
8	Захист МКР	12.12.2019р.

7.2 Графічні матеріали:
- UML діаграма варіантів використання: «06» 12. 2019 р.
- UML-діаграма діяльності: «07» 12. 2019 р..
- UML-діаграма класів: «07» 12. 2019 р.
- ER-діаграма структури бази даних: «08» 12. 2019 р.
- Об’єкт, мета та предмет дослідження. «09» 12. 2019 р.
- Завдання дослідження. «09» 12. 2019 р.
- Приклад категоризації (1 Етап). «10» 12. 2019 р.
- Приклад категоризації (2 Етап). «10» 12. 2019 р.
- Схожість наукових спеціальностей у системі ANZSRC. «10» 12. 2019 р.

8. Порядок контролю і приймання.

8.1. Хід виконання роботи контролюється керівником роботи. Рубіжний контроль провести до «7» грудня 2019 р.

8.2. Атестація проекту здійснюється на попередньому захисті. Попередній захист магістерської кваліфікаційної роботи провести до «10» грудня 2019 р.

8.3. Підсумкове рішення щодо оцінки якості виконання роботи приймається на засіданні ЕК. Захист магістерської кваліфікаційної роботи провести до «17» грудня 2019 р.

Додаток Б Лістинг програм
Лістинг програми визначення схожості спеціальностей

```
<?php

namespace Dimensions\Science\Similarity;



use Common\Science\Similarity\ScienceSimilarity;

use WebParser\Common\ArrayDataRequest;

use WebParser\Common\InMemoryCacheDataSource;

use WebParser\Common\KeyValueRequest;


class DimensionsScienceSimilarity implements ScienceSimilarity

{


  /**

   * @var InMemoryCacheDataSource

   */

  private $dimensionsDataSource;


  /**

   * @var KeyValueRequest[]

   */

  private $years;


  /**

   * DimensionsScienceSimilarity constructor.

   * @param InMemoryCacheDataSource $dimensionsDataSource

   * @param KeyValueRequest[] $years

   */

  public function __construct(InMemoryCacheDataSource $dimensionsDataSource, array $years)

  {

    $this->dimensionsDataSource = $dimensionsDataSource;

    $this->years = $years;

  }


  /**

   * Measures the similarity between two sciences.

   *

   * @param string $scienceName1

   * @param string $scienceName2

   * @return double

   */

  function measure($scienceName1, $scienceName2)

  {

    $fields = $this->dimensionsDataSource->provideData(new ArrayDataRequest(array_merge($this->years, array(new KeyValueRequest("fileName", "AllFields")))));

    $firstId = 0;

    $secondId = 0;

    $firstPublications = 0;

    $secondPublications = 0;

    foreach ($fields as $field) {

      if (strcmp(strtolower($scienceName1), strtolower($field->getName()->getLabel())) == 0) {

        $firstId = $field->getId();

        $firstPublications = $field->getCount()->getLabel();
```

```
        } else if (strcmp(strtolower($scienceName2), strtolower($field->getName()->getLabel())) == 0) {
            $secondId = $field->getId();
            $secondPublications = $field->getCount()->getLabel();
        }
    }

    if ($firstId == 0 || $secondId == 0) {
        throw new \InvalidArgumentException("Could not find either $scienceName1 or $scienceName2");
    }

    $secondRequest = array_merge($this->years, array(new KeyValueRequest("and_facet_for", $secondId), new KeyValueRequest("fileName", $secondId)));
    $secondFields = $this->dimensionsDataSource->provideData(new ArrayDataRequest($secondRequest));
    foreach ($secondFields as $secondField) {
        if (strcmp($secondField->getId(), $firstId) == 0) {
            $intersectionCount = $secondField->getCount()->getLabel();
            return $intersectionCount / ($firstPublications + $secondPublications - $intersectionCount);
        }
    }

    return 0.0;
  }
}
```

Продовження додатку Б

Лістинг програми отримання даних з бази публікацій

```
<?php

namespace WebParser\Dimensions;



use Common\Serializer\Serializer;

use Monolog\Logger;

use WebParser\Common\DataRequest;

use WebParser\Common\NetworkDataSource;


class DimensionsFieldsDataSource extends NetworkDataSource

{


   private static $citationsUrl =
"https://app.dimensions.ai/viz/data/publication/for/aggregated-
citations.json?viz-
st:aggr=mean&search_type=kws&search_field=full_search";


   /**

    * @var Serializer

    */

   private $serializer;


   /**

    * @var Logger

    */

   private $logger;


   /**

    * DimensionsFieldsDataSource constructor.

    * @param Serializer $serializer

    * @param Logger $logger

    */

   public function __construct(Serializer $serializer, Logger $logger)

   {

      $this->serializer = $serializer;

      $this->logger = $logger;

   }


   /**

    * @param DataRequest $request

    * @return DimensionsJsonResult|DimensionsJsonResult[]|null

    */

   function provideData(DataRequest $request)

   {

      $getParams = $request->request();

      $query = "";

      foreach ($getParams as $v) {

         if (strcmp("fileName", $v->getKey()) != 0) {

            $query .= "&" . $v->getKey() . "=" . urlencode($v->getValue());

         }

      }

      $url = self::$citationsUrl . $query;

      return $this->parseByUrl($url);

   }


   /**

    * @param string $url

    * @return DimensionsJsonResult[]

    */

   private function parseByUrl($url)

   {

      $json = $this->get($url);

      $this->logger->debug($url);

      $dimensionsResponse = $this->serializer->deserialize($json,
DimensionsJsonResponse::class);

      $results = $dimensionsResponse->getResults();

      $hasMore = $dimensionsResponse->getNavigation()->getMore()-
>getJson() !== null;

      $page = 1;
```

```
        while ($hasMore) {

            sleep(3); // to not get blocked by server

            $pageUrl = $url . "&page=$page";

            $json = $this->get($pageUrl);

            $this->logger->debug($pageUrl);

            $dimensionsResponse = $this->serializer->deserialize($json,
DimensionsJsonResponse::class);

            if ($dimensionsResponse != null && $dimensionsResponse-
>getResults() != null) {

                $results = array_merge($results, $dimensionsResponse-
>getResults());

                $hasMore = $dimensionsResponse->getNavigation()-
>getMore()->getJson() !== null;

                $page += 1;

            } else $hasMore = false;

        }

        return $results;

    }


}
```

Додаток В
(Обов'язковий)

ЗАТВЕРДЖУЮ
Завідувач кафедри КСУ
д.т.н., проф. В.М. Дубовой
_____
« _____» _____2019 р.

ПЕРЕЛІК
ГРАФІЧНИХ МАТЕРІАЛІВ

для захисту магістерської кваліфікаційної роботи
на тему

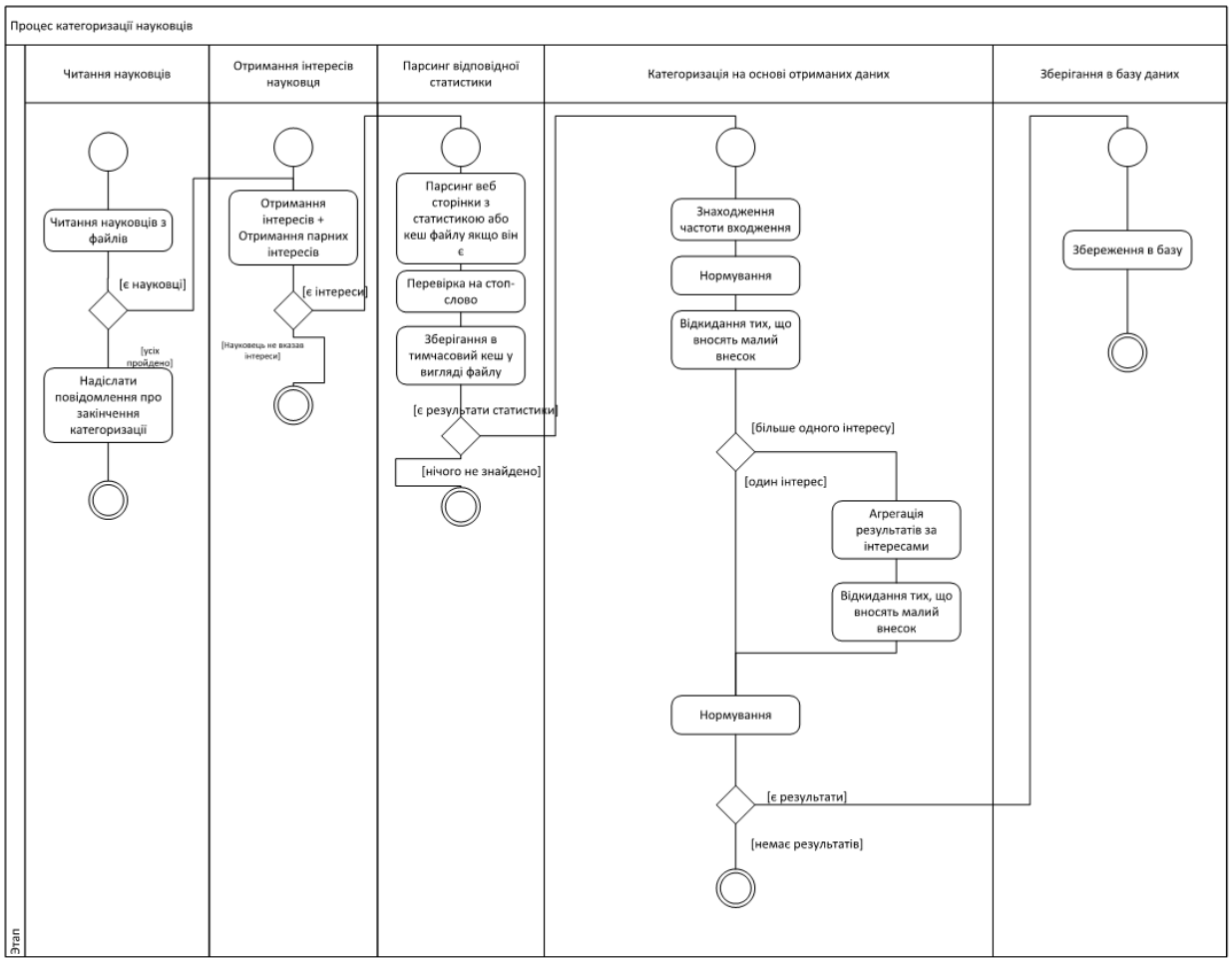
ІНФОРМАЦІЙНА СИСТЕМА ІЄРАРХІЧНОЇ КАТЕГОРИЗАЦІЇ НАУКОВЦІВ ЗА ВІДКРИТИМИ ДАНИМИ З GOOGLE SCHOLAR. ЧАСТИНА 1. ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ

1. UML-діаграма варіантів використання.
2. UML-діаграма діяльності.
3. UML-діаграма класів.
4. ER-діаграма структури бази даних.
5. Таблиця схожості перших 30 спеціальностей.
6. Об'єкт, мета та предмет дослідження.
7. Завдання дослідження.
8. Приклад категоризації (1 Етап).
9. Приклад категоризації (2 Етап).
10. Схожість наукових спеціальностей у системі ANZSRC.

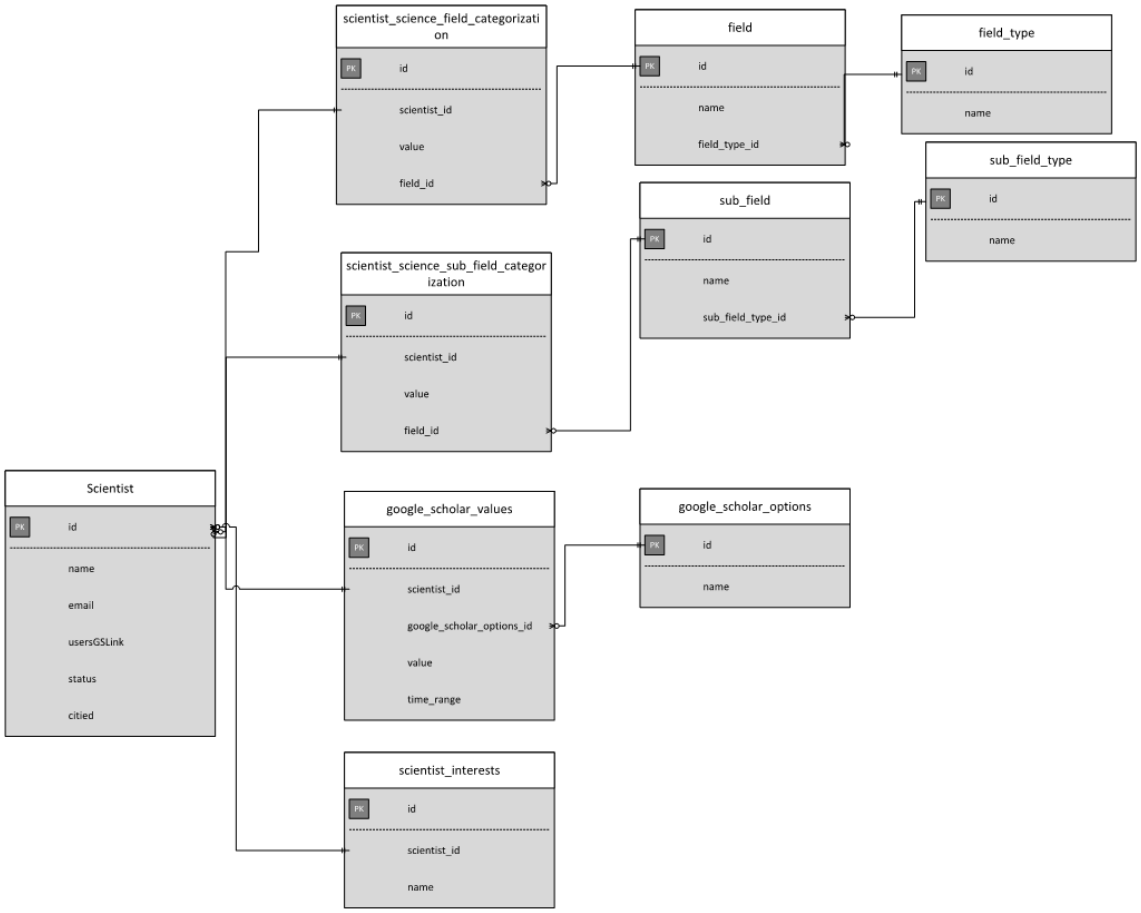
Розробив: Петричко М. В.
_____ _____
(підпис) (дата)
Перевірив: Штовба С. Д.
_____ _____
(підпис) (дата)
Рецензент: Софина О. Ю.
_____ _____
(підпис) (дата)

Вінниця 2019

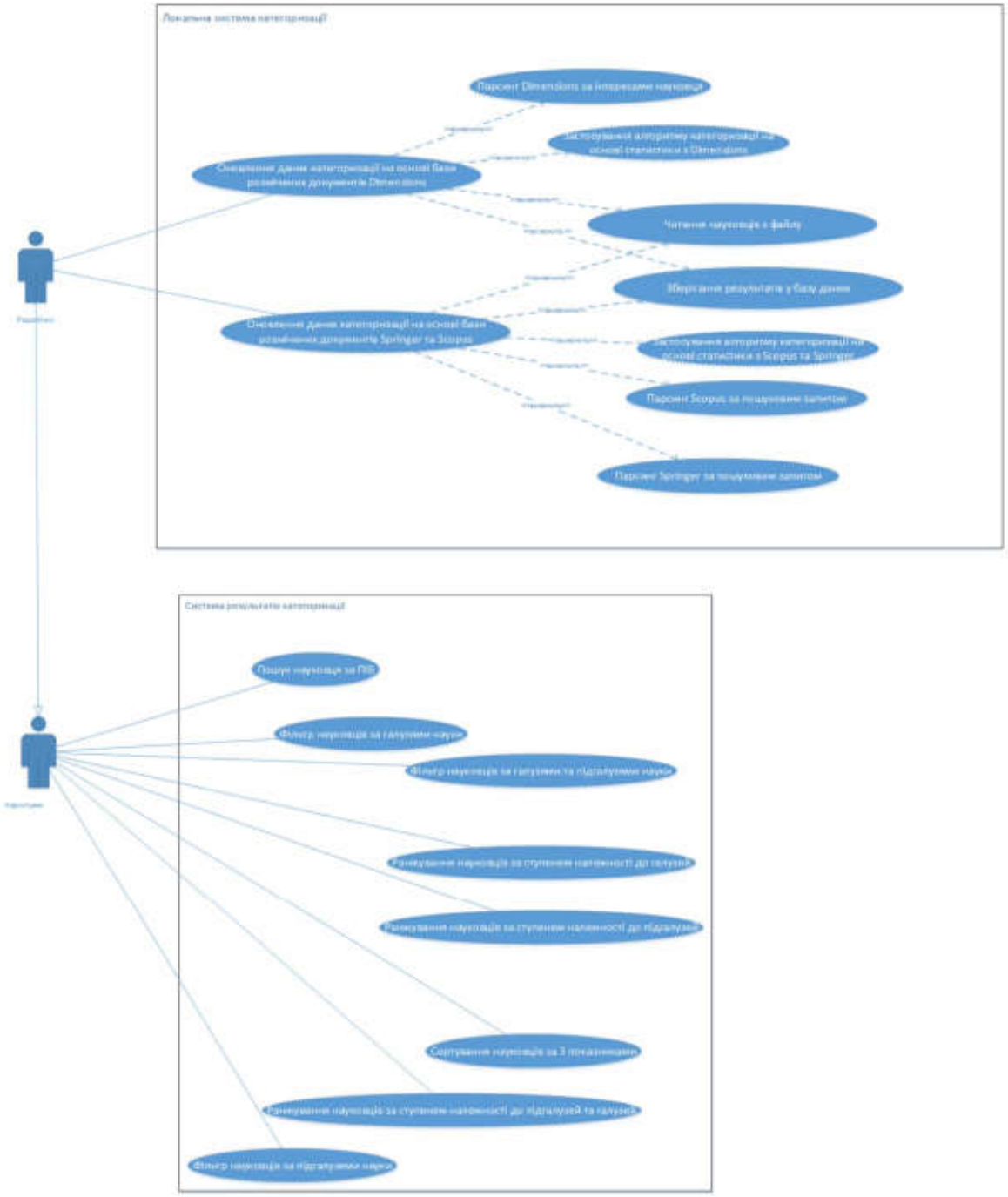
UML-діаграма діяльності алгоритму категоризації



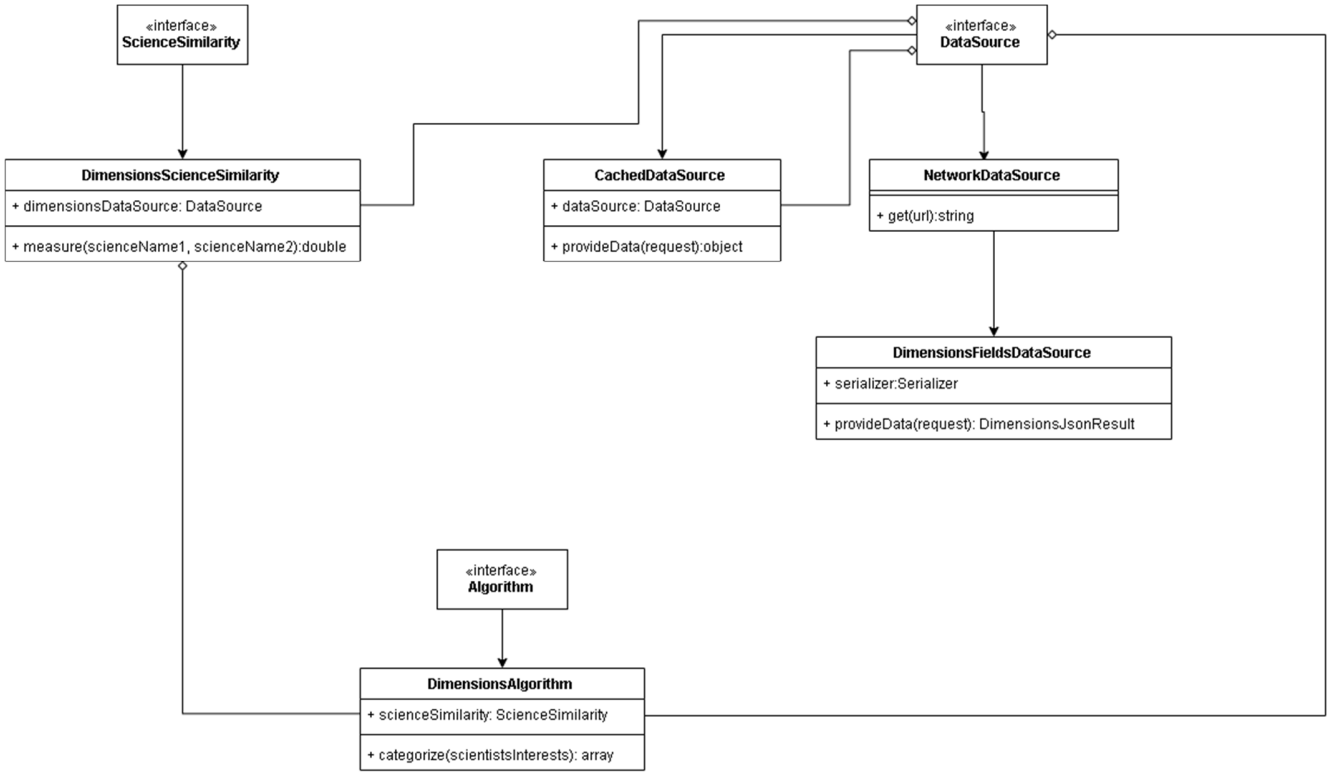
ER-діаграма структури бази даних



UML-діаграма варіантів використання



UML-діаграма частини модулю категоризації науковців



Перших 30 пар найбільш схожих наукових спеціальностей за класифікацією наук ANZSRC

Specialist Studies In Education	Curriculum and Pedagogy	0,258338077
Urban and Regional Planning	Transportation and Freight Services	0,256449932
Art Theory and Criticism	Visual Arts and Crafts	0,240287959
Tourism	Commercial Services	0,223618502
Communication and Media Studies	Journalism and Professional Writing	0,19993908
Ecology	Environmental Science and Management	0,174262271
Organic Chemistry	Medicinal and Biomolecular Chemistry	0,169355979
Geology	Geochemistry	0,167642536
Linguistics	Language Studies	0,149601424
Communication and Media Studies	Film, Television and Digital Media	0,139771264
Food Sciences	Animal Production	0,134035715
Marketing	Commercial Services	0,123572075
Zoology	Evolutionary Biology	0,12282528
Marketing	Tourism	0,119675497
Industrial Biotechnology	Environmental Biotechnology	0,113914208
Human Geography	Urban and Regional Planning	0,112833086
Applied Mathematics	Numerical and Computational Mathematics	0,111342588
Applied Economics	Econometrics	0,10442045
Political Science	Policy and Administration	0,100307506
Psychology	Cognitive Sciences	0,098377847
Atomic, Molecular, Nuclear, Particle and Plasma Physics	Astronomical and Space Sciences	0,096268849
Banking, Finance and Investment	Accounting, Auditing and Accountability	0,095131962
Cultural Studies	Literary Studies	0,093566489
Building	Architecture	0,093324868
Analytical Chemistry	Other Chemical Sciences	0,091489454
Electrical and Electronic Engineering	Communications Technologies	0,088956066
Historical Studies	Literary Studies	0,085845703
Geology	Geophysics	0,085621042
Architecture	Design Practice and Management	0,08545372
Specialist Studies In Education	Curriculum and Pedagogy	0,258338077

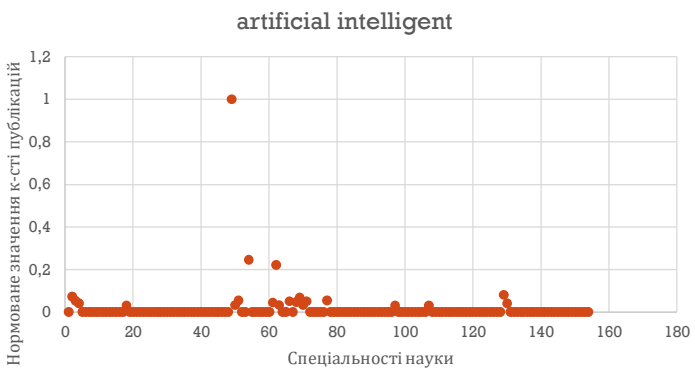
ОБ'ЄКТ, ПРЕДМЕТ ТА МЕТА ДОСЛІДЖЕННЯ

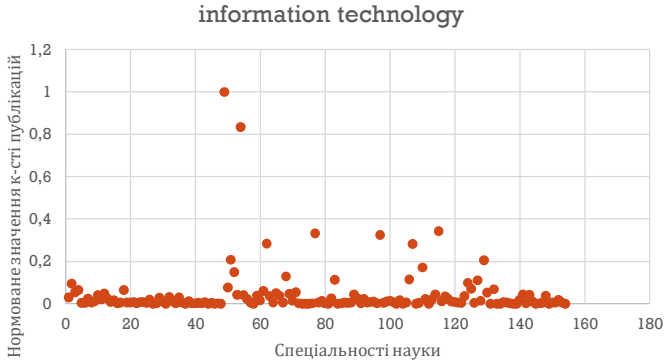
- ***Об'єкт дослідження*** – інформаційні системи підтримки наукової діяльності.
- ***Предмет дослідження*** - моделі та алгоритми ієрархічної категоризації науковців.
- ***Метою роботи*** є розробка інформаційної системи ієрархічної категоризації науковців за відкритими даними з Google Scholar.

ЗАВДАННЯ ДОСЛІДЖЕННЯ

- розробити модель представлення науковців у просторі галузей та спеціальностей за деякої специфікації наук;
- дослідити різноманітні системи класифікації наук та обгрунтувати вибір найбільш придатної для вирішення задачі категоризації;
- дослідити схожість наукових спеціальностей та розробити модель їх оцінювання;
- покращити алгоритм категоризації науковців за спеціальностями який враховує взаємодію як інтересів науковців, так і наукових спеціальностей;
- реалізувати програмно запропоновані моделі та алгоритми.

ПРИКЛАД КАТЕГОРИЗАЦІЇ (І ЕТАП)





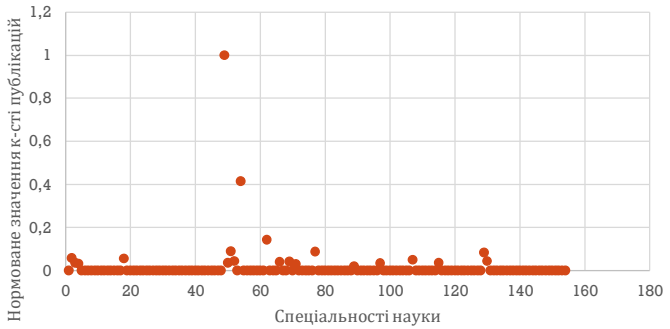
Roman Tkachenko

ПІДПИСАТИСЬ

Dr.Sc, Prof., Lviv Polytechnic National University

Підтверджена електронна адреса в lpnu.ua - Домашня сторінка

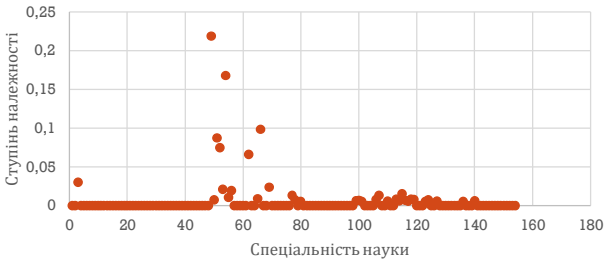
artificial intelligent information technology



2019 ВНТУ

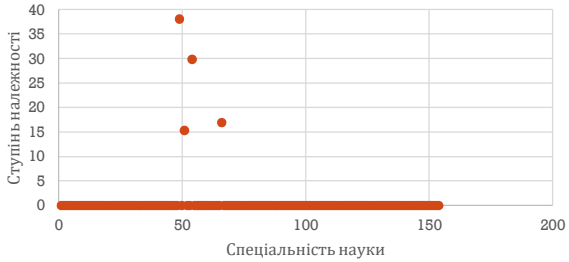
ПРИКЛАД КАТЕГОРИЗАЦІЇ

Ступінь належності після усереднення за інтересами **(II)**



Після нехтування непопулярних спеціальностей **(III)**

Повторне нехтування **(IV)**



Результат категоризації:

- **08 Information and Computing Science – 83.11**
  - 0801 Artificial Intelligence and Image Processing – **37.98**
  - 0806 Information Systems – **29.82**
  - 0803 Computer Software – **15.27**
- **09 Engineering – 16.91**
  - 0910 Manufacturing Engineering – **16.91**

СХОЖІСТЬ НАУКОВИХ СПЕЦІАЛЬНОСТЕЙ У СИСТЕМІ ANZSRC