МАГІСТЕРСЬКА РОБОТА

МР. ІПм - 53.00.00.000 ПЗ

Група ІПм-24-3

Хлібкевич Ярослав

2025

Івано-Франківський національний технічний університет нафти і газу

Факультет інформаційних технологій

Кафедра інженерії програмного забезпечення

Хлібкевич Ярослав Андрійович

(прізвище, ім'я, по батькові)

УДК 004.9

(індекс)

МАГІСТЕРСЬКА РОБОТА

Моделі, методи та засоби побудови рекомендаційної системи відбору фахових наукових конференцій відповідно до профілю науковця

(назва роботи)

Інженерія програмного забезпечення

(назва освітньої програми)

121 - Інженерія програмного забезпечення

(шифр і назва спеціальності)

Хлібкевич Я.А.

(підпис, ініціали та прізвище здобувача освітнього ступеня)

Науковий керівник Корнута Володимир Андрійович, к.т.н., доцент

(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Допущено до захисту

Завідувач кафедри

доц. Бандура В.В.

(посада) (підпис) (дата) (ініціали та прізвище)

Нормоконтроль

доц. Вовк Р.Б.

(посада) (підпис) (дата) (ініціали та прізвище)

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

Івано-Франківськ – 2025

Івано-Франківський національний технічний університет нафти і газу

Факультет інформаційних технологій

Кафедра інженерії програмного забезпечення

Освітній рівень магістр

Спеціальність 121 – Інженерія програмного забезпечення

ЗАТВЕРДЖУЮ:
Зав. кафедрою **ІПЗ**

доц. В.В. Бандура

“ 04 ” вересня 2025 р.

ЗАВДАННЯ

НА МАГІСТЕРСЬКУ РОБОТУ СТУДЕНТУ

Хлібкевичу Ярославу Андрійовичу

(прізвище, ім’я, по-батькові)

1. Тема магістерської роботи “ Моделі, методи та засоби побудови рекомендаційної системи відбору фахових наукових конференцій відповідно до профілю науковця ”

керівник проекту (роботи) Корнута В.А., к.т.н., доцент

затверджені наказом закладу вищої освіти від “ 05 ” листопада 2025 р. № 695/7

2. Строк подання студентом проекту (роботи) 15 грудня 2025 р.

3. Вихідні дані до проекту (роботи) Формальні моделі і методи побудови інформаційних та програмних технологій функціонування рекомендаційних систем

4. Зміст розрахунково - пояснювальної записки(перелік питань, які потрібно розробити)

1. Аналіз предметної області проектування рекомендаційної системи відбору фахових публікацій
2. Розробка методології та проектування архітектури рекомендаційної системи
3. Розробка класифікаторів рекомендаційної системи на основі одновимірних нейронних мереж
4. Оцінка та імплементація моделі для побудови рекомендаційної системи відбору конференцій

5. Перелік графічного матеріалу (з точним зазначенням обов’язкових креслень)

1. Архітектура моделі PC, що складається із взаємопов'язаних матриць і тензора (рис. 1.1)
2. Графічне представлення моделі (рис. 1.2)
3. Ілюстрація цитаційної мережі у тривимірному та двовимірному просторах (рис. 1.3)
4. Архітектура рекомендаційної системи PVR (рис. 1.4)
5. Підхід класифікатора окремих платформ (рис. 1.5)

6. Консультанти розділів проекту (роботи)

Розділ	Консультант	Підпис, дата
Перевірка на плагіат	доц., к.т.н. Вовк Р.Б.	

7. Дата видачі завдання 04 вересня 2025 р.

Керівник _____
(підпис)

Завдання прийняв до виконання_____
(підпис)

КАЛЕНДАРНИЙ ПЛАН

№ п/п	Назви етапів магістерської роботи	Строк виконання етапів роботи	Примітка
1	Підбір і вивчення літератури по темі магістерської роботи	15.09.2025	виконано
2	Аналіз предметної області проектування рекомендаційної системи відбору фахових публікацій	01.10.2025	виконано
3	Розробка методології та проектування архітектури рекомендаційної системи	17.10.2025	виконано
4	Розробка класифікаторів рекомендаційної системи на основі одновимірних нейронних мереж	02.11.2025	виконано
5	4. Оцінка та імплементація моделі для побудови рекомендаційної системи відбору конференцій	19.11.2025	виконано
6	Опис пропонованої методології щодо побудови рекомендаційної системи вибору платформи фахових публікацій	02.12.2025	виконано
7	Затвердження пояснювальної записки роботи завідувачем кафедри	15.12.2025	виконано

Студент – магістр _____
(підпис)

Керівник роботи _____
(підпис)

АНОТАЦІЯ

Магістерська робота: 75 с., 19 рис., 9 табл., 41 джерел.

Тема: Моделі, методи та засоби побудови рекомендаційної системи відбору фахових наукових конференцій відповідно до профілю науковця

Метою роботи є розроблення моделей, методів та засобів побудови рекомендаційної системи відбору фахових наукових конференцій на основі методів глибокого навчання та аналізу текстових даних.

Об'єктом дослідження є процес формування рекомендацій щодо вибору платформ наукових конференцій.

Предметом дослідження є методи, моделі та архітектурні рішення побудови рекомендаційної системи відбору конференцій із використанням методів глибокого навчання та текстової аналітики.

Результати дослідження

В роботі сформовано теоретичну основу побудови рекомендаційних моделей у контексті наукових публікацій, розроблено архітектуру модульної рекомендаційної системи, що базується на методах глибокого навчання та текстової аналітики.

Висновок

Обґрунтовано актуальність розробки рекомендаційної системи для відбору фахових наукових платформ, що відповідає зростаючому обсягу наукової інформації та потребам персоналізації та визначено оптимальні підходи до підвищення продуктивності системи, включаючи балансування наборів даних та ансамблеві класифікатори.

РЕКОМЕНДАЦІЙНА СИСТЕМА, НАУКОВІ КОНФЕРЕНЦІЇ, ГЛИБОКЕ НАВЧАННЯ, ОБРОБКА ПРИРОДНОЇ МОВИ, КЛАСИФІКАЦІЯ, ТЕКСТОВІ ДАНІ, ПОДІБНІСТЬ ДОКУМЕНТІВ, НЕЙРОННІ МЕРЕЖІ

ABSTRACT

Master Thesis: 75 pp., 19 fig., 9 tab., 41 sources.

Topic: Models, methods and tools for building a recommendation system for selecting professional scientific conferences according to the profile of a scientist

The purpose of the work is to develop models, methods and tools for building a recommendation system for selecting professional scientific conferences based on deep learning methods and text data analysis.

The object of the research is the process of forming recommendations for choosing scientific conference platforms.

The subject of the research is methods, models and architectural solutions for building a recommendation system for selecting conferences using deep learning methods and text analytics.

Research results

The work forms a theoretical basis for building recommendation models in the context of scientific publications, and develops the architecture of a modular recommendation system based on deep learning and text analytics methods.

Conclusion

The relevance of developing a recommender system for selecting professional scientific platforms that meets the growing volume of scientific information and personalization needs is substantiated, and optimal approaches to increasing the system's performance are identified, including data set balancing and ensemble classifiers.

RECOMMENDATION SYSTEM, SCIENTIFIC CONFERENCES, DEEP LEARNING, NATURAL LANGUAGE PROCESSING, CLASSIFICATION, TEXT DATA, DOCUMENT SIMILARITY, NEURAL NETWORKS

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

РС – рекомендаційна система

CNN - Convolutional Neural Networks

DL - Digital Library

ACM - Association for Computing Machinery

JCDL - Joint Conference on Digital Libraries

ANN - Artificial Neural Networks

SIG Special Interest Group

CCS - Computing Classification System

POS - Part-Of-Speech

MCI - Mild Cognitive Impairment

CN - Cognitive Normal

HPC - High Performance Computing

ВСТУП

Актуальність теми.

Сучасний науковий простір характеризується стрімким зростанням кількості наукових публікацій, конференцій, журналів та комунікаційних платформ, що формують високодинамічне середовище для обміну знаннями. У таких умовах науковці стикаються з необхідністю оперативного та обґрунтованого вибору конференцій, які найбільш точно відповідають їхній дослідницькій спеціалізації, рівню підготовки та науковій тематиці. Невідповідність між тематикою дослідження та профілем конференції може призвести до низької ефективності презентації результатів, втрати наукової аудиторії та уповільнення професійного розвитку. Водночас кількість доступних міжнародних конференцій зростає, а разом із нею ускладнюється процес індивідуального добору релевантних платформ.

У цьому контексті рекомендаційні системи стають необхідним інструментом підтримки прийняття рішень у науковій діяльності. Проте більшість існуючих систем не враховують індивідуальний профіль науковця та ґрунтуються переважно на загальних метриках подібності, що знижує точність рекомендацій. Крім того, процес відбору конференцій часто потребує аналізу великого обсягу текстової інформації, яка включає анотації, ключові слова, тематичні описи та історію публікацій. Це створює потребу у використанні сучасних методів обробки природної мови та глибокого навчання, здатних автоматично інтерпретувати зміст текстових документів і визначати їхню релевантність.

Розробка рекомендаційної системи, що базується на глибинних нейронних мережах і методах текстової аналітики, є важливим кроком до створення інтелектуальних інструментів для наукового середовища. Така система дозволить не лише автоматизувати процес добору конференцій, а й забезпечити високу точність класифікації наукових профілів і платформ. Дана магістерська робота присвячена розробці методів, моделей та технічних

засобів побудови рекомендаційної системи нового покоління, що дозволяє підвищити ефективність прийняття рішень щодо участі у фахових наукових конференціях.

Актуальність дослідження зумовлена зростаючою потребою науковців у швидкому й точному визначенні конференцій, що відповідають їхнім дослідницьким інтересам. У глобальному інформаційному просторі кількість наукових подій щороку збільшується, а їх тематика постійно розширюється, що робить процес самостійного аналізу платформ надзвичайно трудомістким. Традиційні підходи до відбору конференцій не відповідають сучасним вимогам, оскільки спираються на ручний перегляд профілів конференцій, не враховують особливості наукового стилю автора та мають низький рівень автоматизації.

Паралельно з цим зростає роль рекомендаційних систем, однак більшість із них не охоплюють специфіку наукової діяльності, де ключове значення має якість і точність текстового аналізу. Розвиток методів машинного навчання та глибоких нейронних мереж створює нові можливості для побудови інтелектуальних систем, які здатні обробляти складні текстові структури та знаходити приховані семантичні зв’язки. Водночас наявні дослідження демонструють брак систем, які здатні інтегрувати класифікаційні моделі та моделі подібності в єдину комплексну рекомендаційну технологію. Саме тому створення системи відбору наукових конференцій на основі глибокого навчання є своєчасним, науково обґрунтованим і практично значущим завданням.

Метою роботи є розроблення моделей, методів та засобів побудови рекомендаційної системи відбору фахових наукових конференцій на основі методів глибокого навчання та аналізу текстових даних.

Об’єктом дослідження є процес формування рекомендацій щодо вибору платформ наукових конференцій.

Предметом дослідження є методи, моделі та архітектурні рішення побудови рекомендаційної системи відбору конференцій із використанням методів глибокого навчання та текстової аналітики.

Завдання дослідження:

1. Проаналізувати предметну область рекомендаційних систем і визначити вимоги до системи для відбору конференцій.

2. Дослідити сучасні методи рекомендації, класифікації та аналізу текстових даних у контексті наукових платформ.

3. Розробити методологію побудови системи, включаючи архітектуру, моделі та компоненти обробки даних.

4. Реалізувати класифікаційні моделі на основі глибоких нейронних мереж та оцінити їхню ефективність.

5. Створити та протестувати прототипи рекомендаційної системи.

Методи дослідження

У роботі використано методи аналізу та синтезу наукової інформації, методи обробки природної мови, техніки токенізації та векторизації текстів, методи глибокого навчання, зокрема одновимірні згорткові нейронні мережі, методи балансування даних, а також метрики семантичної схожості. Для оцінки якості використовувались статистичні методи класифікації, метрики точності, повноти та F1-міри.

Наукова новизна отриманих результатів

Наукова новизна роботи полягає у розробленні комплексної методології побудови рекомендаційної системи, яка поєднує класифікаційні моделі глибокого навчання з методами семантичної подібності між платформами. Синтезовано механізм групових класифікаторів для підвищення стабільності рекомендацій у сфері вибору наукових конференцій. Запропоновано архітектуру системи, яка забезпечує модульність і масштабованість та дозволяє поєднувати різні підходи до аналізу текстових ознак.

Практичне застосування результатів

Результати роботи можуть бути використані для створення практичної рекомендаційної системи для університетів, наукових установ, дослідницьких груп та індивідуальних науковців. Розроблені моделі та архітектурні рішення можуть бути інтегровані у платформи наукометричного аналізу, системи управління науковими проектами та інформаційні наукові портали.

Структура магістерської роботи. Представлена робота складається зі вступу, трьох розділів та висновків. Загальний обсяг роботи становить 75 сторінок, і містить 19 рисунків, 9 таблиць, перелік використаних джерел із 41 найменування.

РОЗДІЛ 1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ПРОЕКТУВАННЯ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ ВІДБОРУ ФАХОВИХ ПУБЛІКАЦІЙ

1.1. Контентно-орієнтований підхід до побудови рекомендаційної системи вибору публікаційних платформ

Основною метою дослідницької діяльності є дисемінація наукових результатів шляхом публікації у формі статей у рецензованих журналах (періодичних виданнях) та матеріалах визнаних наукових конференцій. Існує ключова відмінність у форматі цих наукових майданчиків: конференції, як правило, накладають суворі обмеження на обсяг (кількість сторінок) поданих матеріалів, тоді як журнали часто демонструють більшу гнучкість щодо ліміту слів чи сторінок. Щодо періодичності, академічні конференції переважно мають щорічний цикл проведення, тоді як журнали функціонують із регулярними термінами подання (щомісячними, щоквартальними тощо).

У контексті значної диверсифікації наукових напрямків і мікрогалузей, що супроводжується зростанням кількості потенційних платформ для публікації (які також включають семінари, симпозіуми та конкурси), завдання оптимального вибору відповідного веню для подання рукопису стає нетривіальним і складним завданням. Неадекватний вибір публікаційного майданчика може призвести до відхилення роботи. Таким чином, перед кожним автором постає критичне питання, що потребує систематичного вирішення: "Яким чином ідентифікувати найбільш релевантне та перспективне місце для публікації з метою максимізації ймовірності прийняття рукопису?"

Дана магістерська робота присвячена розробці системи рекомендацій (СР), яка слугуватиме інструментом для вирішення вищезазначеної проблеми. СР є визнаним інструментом підтримки прийняття рішень, що застосовується у багатьох доменах. Класичним прикладом є механізм фільтрації на основі спільної поведінки (наприклад, "Клієнти, які придбали

[товар X], також придбали [товар Y]"), що широко використовується у сфері електронної комерції для полегшення навігації користувачів у великих каталогах і задоволення їхніх інформаційних потреб.

Метою цього дослідження є створення спеціалізованої системи рекомендацій, призначеної для допомоги науковцям у виборі платформи для публікації.

Процес подання наукової роботи передбачає структурування рукопису, що містить: назву дослідження, реферативне резюме (анотацію), перелік відповідних ключових слів та детальний опис виконаної роботи. Розроблена система здатна обробляти будь-який з цих компонентів як вхідні дані та генерувати перелік релевантних публікаційних майданчиків на основі контентного аналізу поданого матеріалу.

Для реалізації проекту було використано набір метаданих, отриманий із цифрової бібліотеки ACM (Association for Computing Machinery). Рекомендаційний механізм було розроблено на основі методів глибокого навчання (Deep Learning). Очікується, що функціональність системи забезпечить можливість ідентифікації найбільш оптимальної платформи або групи високорелевантних публікаційних платформ для дослідницької роботи.

1.2. Опис актуальності розробки рекомендаційної системи для вибору найбільш релевантних платформ для наукових публікацій

Академічна дослідницька діяльність охоплює низку ключових процесів, включаючи дисемінацію результатів шляхом публікації наукових праць, цитування релевантних джерел та налагодження наукової співпраці. Кожен із цих процесів вимагає прийняття критичних стратегічних рішень, що істотно впливають на результативність досліджень. Дослідники зобов'язані здійснювати обґрунтований вибір предметної галузі, партнерів для кооперації, а також оптимальної наукової платформи для опублікування своїх напрацювань.

У межах даної магістерської роботи зосереджено увагу на проблемі ідентифікації найбільш придатної наукової платформи для подання академічних продуктів (статей, доповідей тощо). Ця дослідницька задача формалізується як "проблема відповідності потенційної наукової платформи" (Platform Matching Problem). Запропоноване рішення має на меті допомогти авторам у виборі найрелевантнішої конференції або журналу для подання рукопису.

Актуальність дослідження обумовлена експоненційним зростанням кількості академічних наукових платформ у сучасному науковому просторі. Окрім підвищення ймовірності наукового визнання, вибір адекватної платформи для публікації також сприяє отриманню високоякісних рецензій, що є критично важливим для подальшого вдосконалення дослідницької роботи.

Кількісні показники ілюструють складність вибору: поточна кількість конференцій ACM (Association for Computing Machinery) становить приблизно 120, не враховуючи конференцій спеціальних груп інтересів [1]. Обсяг публікаційних можливостей у вузьких обчислювальних дисциплінах також є значним. Наприклад, у галузі пошуку інформації функціонує низка впливових конференцій (зокрема, SIGIR, CIKM, WSDM, WWW, JCDL та інші). Пошук на веб-сайті IEEE [6] також демонструє великий перелік організованих асоціацією конференцій. Отже, завдання ідентифікації оптимального набору наукових платформ набуває виняткової значущості.

Розв'язання цієї проблеми безпосередньо корелює з ширшою проблемою інформаційного перевантаження (Information Overload). У ситуаціях, коли кількість потенційних рішень є надмірною, застосування рекомендаційних систем може істотно знизити когнітивне навантаження на процес прийняття рішень. Відомі комерційні корпорації, як-от Amazon та Netflix, успішно використовують персоналізовані рекомендаційні моделі для продуктів, що підвищує рівень задоволеності користувачів та оптимізує доходи. Центральна ідея цієї роботи полягає в трансфері та адаптації

рекомендаційного підходу для вирішення проблеми інформаційного надлишку в академічному середовищі.

Проблема відповідності наукової платформи в академічному секторі є предметом активних досліджень, як засвідчено у роботах [6, 8]. Більше того, провідні видавництва, включаючи Elsevier, IEEE, Springer та Wiley, вже інтегрували рекомендаційні сервіси для підбору відповідних журналів для подання статей. Проте, наявність спеціалізованої рекомендаційної системи для наукових платформ ACM не зафіксована. Ця робота спрямована на усунення цього прогалу, пропонуючи розробку рекомендаційної системи на основі глибокого навчання, оптимізованої для набору даних ACM.

Глибоке навчання (Deep Learning) набуло стрімкої популярності, особливо в сферах комп'ютерного зору та обробки природної мови (NLP). Ефективність класичних рекомендаційних підходів (базованих на контентній фільтрації, колаборативній фільтрації чи їх гібридизації) критично залежить від якості та релевантності ручно-сконструйованих ознак. Трудомістке інженерне конструювання ознак є значним накладним елементом, особливо коли якісні ознаки важко ідентифікувати в наборі даних. Глибоке навчання пропонує перевагу завдяки своїй високій здатності автономно вивчати ієрархічні ознаки (feature learning) безпосередньо з вихідних даних, що робить його найбільш бажаним та ефективним архітектурним рішенням для імплементації рекомендаційних систем [9].

1.3. Огляд літератури щодо методів рекомендації наукових платформ

1.3.1. Еволюція рекомендаційних систем та їх застосування

Рекомендаційні системи (РС) еволюціонували від допоміжної функції до очікуваного атрибуту у більшості сучасних інформаційних доменів, охоплюючи такі популярні сфери застосування, як спорт, новинні стрічки та

кінематограф. Розширення цієї парадигми передбачає імплементацію РС у сфері академічної діяльності.

Проблема рекомендації в академічному середовищі, зокрема рекомендація наукової літератури, є предметом тривалих досліджень. У роботі [9] ідентифіковано понад 80 різних методологічних підходів, використаних для рекомендації академічної літератури. Проте, важливо зазначити, що з усіх проаналізованих робіт лише невелика частина розглядала задачу рекомендації наукових платформ (конференцій та журналів) як самостійну проблему.

Якість рекомендаційної системи (РС) визначається сукупністю трьох ключових факторів: точність (Accuracy), задоволеність користувача (User Satisfaction) та задоволеність провайдера (Provider Satisfaction).

1. Точність (Accuracy) рекомендаційної системи

Першим і фундаментальним фактором, що визначає якість РС, є її точність, тобто здатність задовольняти інформаційну потребу окремого користувача.

Інформаційні потреби користувачів є гетерогенними і варіюються залежно від:

- Освітнього рівня та попередніх знань.
- Персональних уподобань та дослідницьких цілей.
- Контексту запиту.

Наприклад, один користувач може шукати найновіші статті з ментальних карт, тоді як іншого цікавить перша публікація, що вводить поняття рекомендаційних систем, або найпопулярніші медичні дослідження щодо раку легенів, але лише певною мовою. Елементи, що відповідають цим потребам, вважаються релевантними для користувача.

Відповідно, якісна РС — це та, яка рекомендує найбільш релевантні елементи. Для досягнення цієї мети РС має спочатку ідентифікувати інформаційні потреби своїх користувачів, а потім визначити елементи, які їх задовольняють. Точність відображає ефективність виконання цього завдання:

що більше релевантних та менше нерелевантних елементів вона пропонує, то вищою є її точність.

Обов'язковою умовою досягнення високої точності є високе охоплення наявних елементів [5]. Охоплення описує частку документів у базі даних РС, які в принципі можуть бути рекомендовані за допомогою використаного підходу.

Для текстових підходів охоплення зазвичай становить 100%.

Для багатьох підходів, заснованих на цитуванні, охоплення є значно нижчим, оскільки лише частина всіх документів цитується, і, отже, може бути рекомендована.

2. Задоволеність користувача

Другий фактор, що визначає якість РС, – це її здатність забезпечувати задоволеність користувача. Хоча на перший погляд можна припустити, що точна РС автоматично задовольняє користувача, на задоволеність впливають численні додаткові чинники.

Додаткові чинники задоволеності:

- Серендипність (Serendipity) – хоча точність є важливою, іноді надто очевидна, але точна рекомендація (наприклад, рекомендація молока клієнту в супермаркеті) може не принести задоволення. Користувачі часто очікують більш різноманітних рекомендацій, які все ж таки зберігають певний рівень точності.

- Оперативність та юзабіліті - користувачі можуть бути незадоволені точною РС, якщо час очікування рекомендації є надто тривалим, презентація є непривабливою, маркування рекомендацій є неоптимальним.

- Прозорість та етика - незадоволеність може виникнути, якщо рекомендації надаються виключно з комерційних міркувань.

- Задоволеність може відрізнятися за демографічними показниками; наприклад, старші користувачі, як правило, більш задоволені рекомендаціями, ніж молодші [8].

- Час, який користувач повинен витратити перед отриманням рекомендацій, також впливає на задоволеність. Системи, які автоматично виводять інтереси користувачів, значно зменшують необхідні часові зобов'язання порівняно з тими, що вимагають ручного введення інтересів.

3. Задоволеність провайдера рекомендацій

Третій фактор, що визначає якість рекомендаційної системи, – це її здатність задовольняти провайдера рекомендацій. Хоча зазвичай припускається, що провайдери задоволені, коли задоволені їхні користувачі, це не завжди так.

Однією з ключових цілей провайдерів є зниження операційних витрат, які можуть вимірюватися робочим часом, дисковим простором, обсягом пам'яті, потужністю CPU та трафіком. Таким чином, якісна рекомендаційна система також може бути визначена як та, що може бути розроблена, експлуатована та підтримувана за низької вартості. Інші провайдери, наприклад видавництва, можуть мати на меті отримання прибутку від РС. У цьому випадку видавець може віддати перевагу рекомендації елементів із вищою маржею прибутку, навіть якщо задоволеність користувача є дещо нижчою.

Новинний веб-сайт може мати за мету утримання читачів на своєму ресурсі якомога довше; у цьому випадку РС переважно пропонуватиме довші статті, навіть якщо коротші могли б призвести до вищої задоволеності користувача.

У більшості випадків між трьома вищезазначеними факторами існує компроміс (tradeoff). Наприклад, використання кластеризації суттєво скорочує час виконання, а отже, й витрати, але часто знижує точність [10]. Аналогічно, коли основною метою є отримання доходу, задоволеність користувача може постраждати. Важливо, що задоволеність користувача ніколи не повинна бути надто низькою, оскільки це призведе до повного ігнорування рекомендацій.

1.3.2. Існуючі підходи до рекомендації наукових платформ

Однією з перших робіт у цій сфері є [11], де розроблено рекомендатор наукових платформ на основі колаборативної фільтрації. У цьому дослідженні використано комбінацію стилометричних ознак та аналізу найближчих сусідів (Nearest Neighbor) на даних CiteSeer. Стилометричні ознаки вмісту статей включали такі параметри, як загальна кількість слів, кількість розділів, кількість ілюстрацій тощо. Автори використовували як метадані, так і повний текст для рекомендації платформ, встановлюючи зв'язок між статтями з ідентичними авторами.

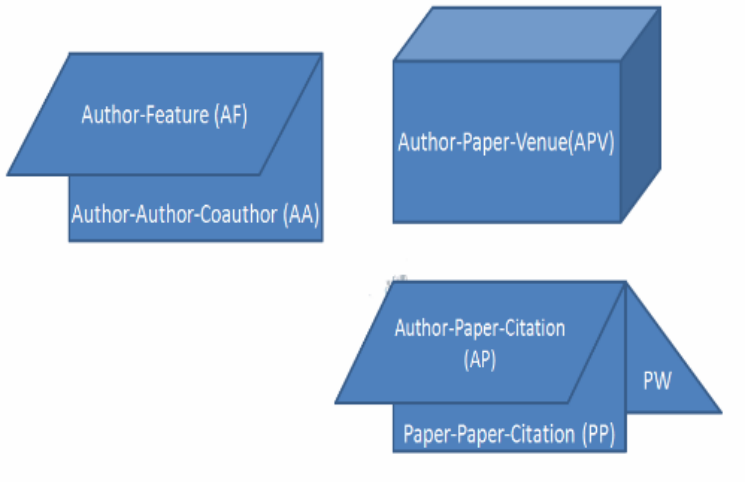

Рис. 1.1. Архітектура моделі рекомендаційної системи, що складається із взаємопов'язаних матриць і тензора

Запропонована модель структурована для вирішення чотирьох специфічних рекомендаційних завдань в академічному середовищі. Кожне завдання відповідає певній академічній активності та індукує одне відношення, що призводить до формування чотирьох основних відношень у моделі:

- Відношення "Автор–Стаття–Наукова платформа" (APV-тензор) - використовується для рекомендації наукових платформ.

- Відношення "Співпраця Автор–Автор" (AA-матриця) - моделює колаборaційні зв'язки між дослідниками.

- Відношення "Автор–Стаття–Цитування" (AP-матриця) - фіксує зв'язок між автором і статтями, які він цитує.

Відношення "Стаття–Стаття–Цитування" (PP-матриця) - моделює прямі цитаційні зв'язки між науковими працями.

Рисунок 1.1. ілюструє загальну архітектуру моделі, що складається із взаємопов'язаних матриць і тензора.

З метою подолання проблеми холодного старту (cold start problem) та підвищення персоналізації для авторів, до моделі додатково інтегруються допоміжні ознаки для статей та авторів. У поточному дослідженні ознаками статті виступає лише чистий контент статті, представлений матрицею "Стаття–Слово" (PW-матриця). Автори та їхні ознаки моделюються за допомогою AF-матриці.

Рисунок 1.2 являє собою графічне представлення цієї моделі, що відображає взаємозв'язки між усіма компонентами.

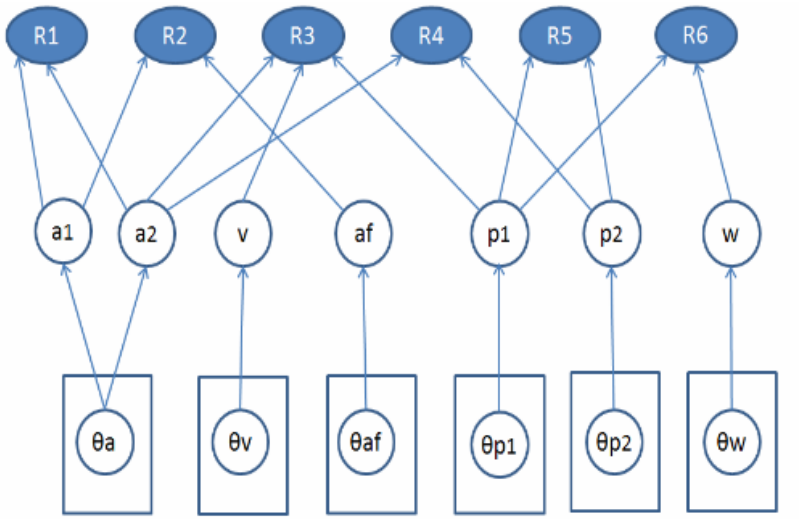

Рис. 1.2. Графічне представлення моделі

Також в роботі формально описуються чотири завдання рекомендації/ прогнозування та вводиться процедура конструювання відповідних відношень.

Інші дослідницькі ініціативи використовували ширший контекст:

Робота [12] застосовує публікаційну мережу автора та контент самого рукопису для прогнозування платформ. Це дослідження використовувало дані конференцій Спеціальних груп інтересів (SIG) ACM.

В дослідженні [7] використовували мережі цитувань конкретного поданого рукопису для формування рекомендацій. Для підвищення результативності аналізу, дана робота концентрується на вивченні схожості академічних об'єктів у межах розширеної цитаційної мережі для заданого запиту. Тут пропонується рекомендаційна модель, яка обчислює коефіцієнт довіри (confidence score) платформ у розширеній мережі, використовуючи бібліографічні ознаки для пропозиції відповідної наукової платформи для подання статті. Автори віртуально розміщують профілі платформ у певному околі (neighborhood), що дозволяє визначити найбільш релевантну платформу.

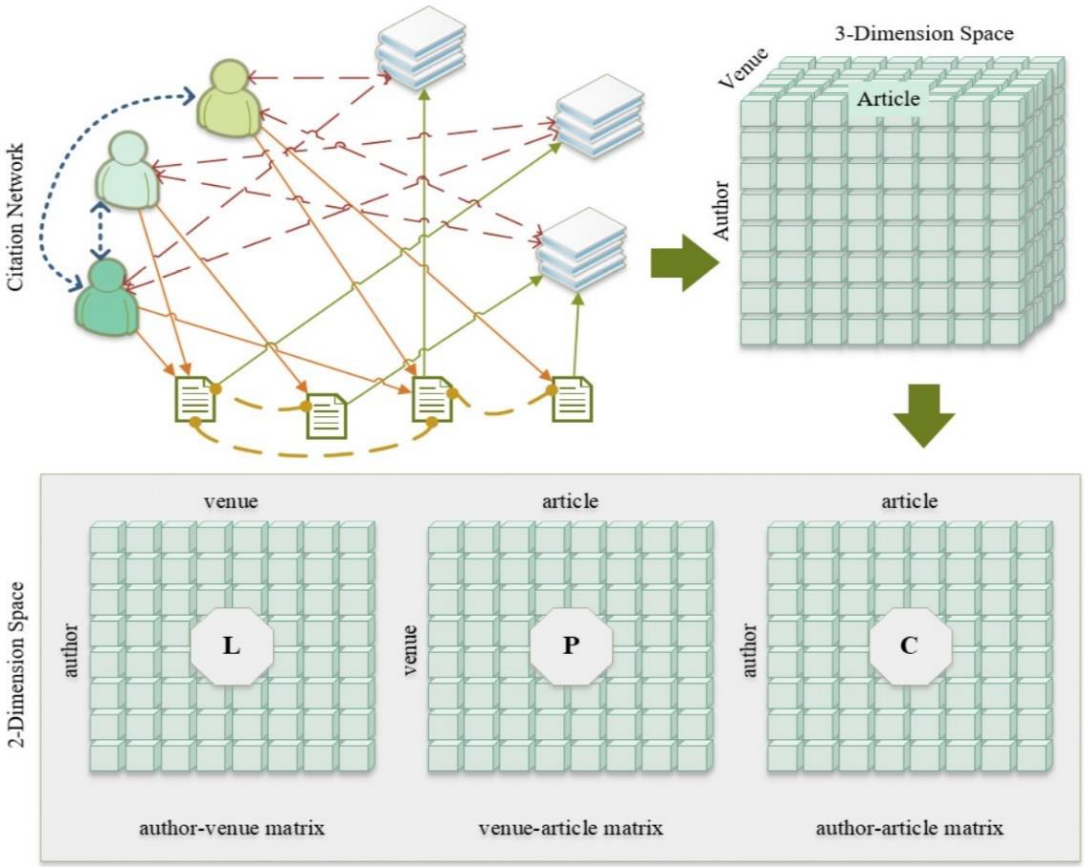

Рис. 1.3. Ілюстрація цитаційної мережі у тривимірному та двовимірному просторах

Запропонована модель спрямована на визначення латентної бажаної асоціації платформи з конкретними статтями та авторами шляхом навчання на основі схожості між академічними об'єктами.

Мотивацією для розробки моделі стала типова поведінка дослідників:

- Вони прагнуть підтримувати контакти з авторами, яких зустрічають на платформах.

- Вони цитують експертів у конкретній дослідницькій галузі.

- Вони шукають високоякісні та успішні наукові платформи.

- Вони беруть участь у конференціях, тісно пов'язаних із їхніми дослідженнями.

- Вони цитують статті з високоякісних платформ та видавництв.

Ця модель є розширенням попередніх робіт, які пропонували академічні рекомендації на основі тематичної схожості, схожості авторів та платформ та продемонстрували значні результати. Авторами пояснюється, як побудувати цитаційну мережу та персоналізувати рекомендаційну модель шляхом виявлення латентних уподобань у цитаційній мережі для вхідного запиту.

Деякі роботи імплементували механізми зворотного зв'язку для підвищення якості пропонованих платформ, а дослідження [10] використовувало інформацію про співавторів, співцитувальників та спільну афіліацію для персоналізації рекомендацій.

Усі вищезгадані підходи застосовують повний текст, цитування та метадані, отримані з різних джерел. На противагу цьому, автори [13] запропонували методи, які надають попередні рекомендації, використовуючи лише назву та анотацію рукопису. У цьому підході для кожної статті під час фази навчання створюється мовна модель на основі n-грам (символьних грам) за методом Кавнара-Тренкла (Cavnar Trenkle). Під час фактичної рекомендації статті пропонуються платформи, чиї мовні моделі мають найкоротшу відстань до тестової статті. Крім того, автори залучали Латентне розміщення Діріхле (Latent Dirichlet Allocation, LDA) для ідентифікації

тематики платформ та тестових статей, де платформи, близькі за евклідовою відстанню в просторі тем, ставали кандидатами на рекомендацію.

Серед застосованих методів, дослідники використовували аналіз тем (через LDA), колаборативну фільтрацію на основі контентних ознак та класифікатори найближчих сусідів. У [11] також застосовувалася матрична/тензорна факторизація. У подібному ключі, робота [15] продемонструвала підхід контентної фільтрації для рекомендації наукових платформ.

Існує також підхід, орієнтований на рекомендацію платформ на основі поточних дослідницьких інтересів користувача [7]. Аналіз наукової читацької поведінки дослідника слугує формою неявного зворотного зв'язку, що, зокрема, допомагає вирішити проблему холодного старту (cold start problem).

Рисунок 1.4 ілюструє архітектуру рекомендаційної системи PVR (Paper-Venue-Researcher).

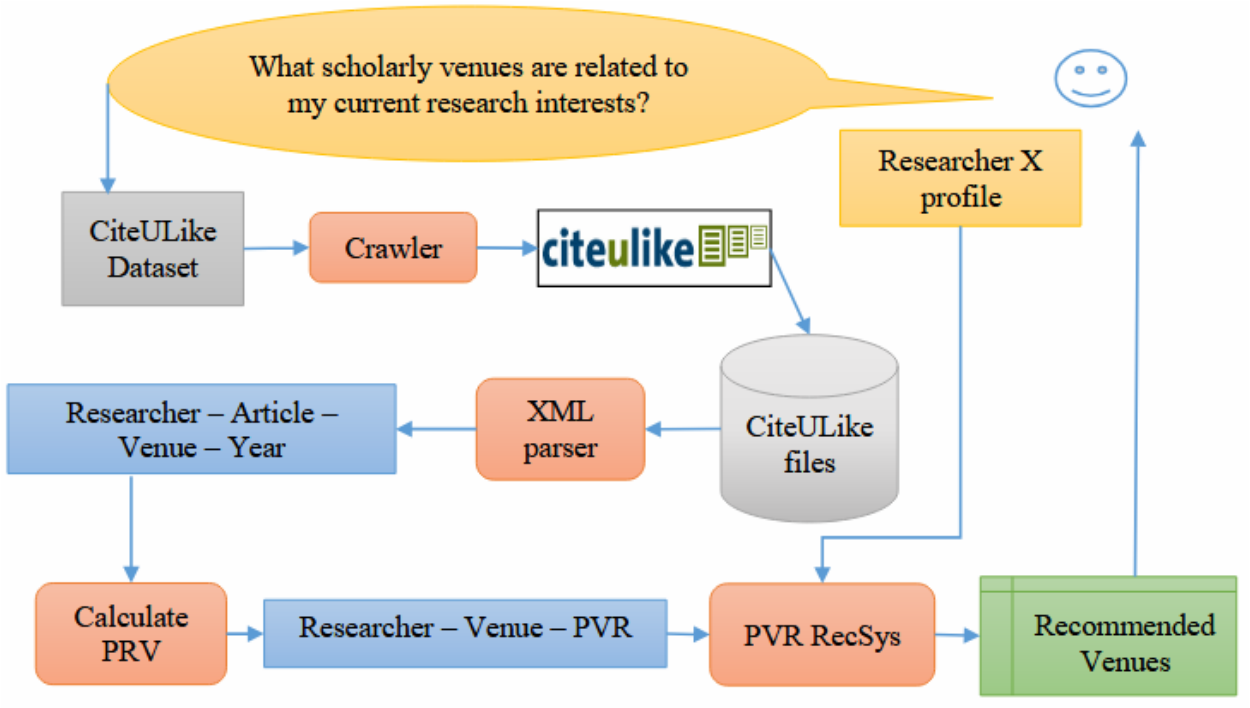

Рис. 1.4. Архітектура рекомендаційної системи PVR

Завданням системи є відповісти на питання: "Які наукові платформи відповідають моїм поточним дослідницьким інтересам?" Система обробляє

вхідні дані про дослідника та публікації, обчислює показники PRV (Paper-Venue-Researcher), а потім використовує ці дані у рекомендаційній системі PVR для надання списку рекомендованих платформ.

Для рекомендації наукових платформ дослідникам було імплементовано наступні алгоритми з пакету Apache Mahout:

- Колаборативна фільтрація, орієнтована на користувача (User-based CF).

- Колаборативна фільтрація, орієнтована на елемент (Item-based CF).

- Стохастичний градієнтний спуск (Stochastic Gradient Descent, SGD).

- Факторизація матриць за допомогою сингулярного розкладу (SVD++).

В експерименті порівнювалися дослідники зі схожими інтересами на основі їхніх показників PVR. У рекомендаційних системах ключовим етапом є визначення схожості між користувачами або елементами.

Також тут астосували три широко використовувані метрики схожості:

1. Косинусна схожість (Cosine Similarity) - визначається як косинус кута між двома векторами. Ця метрика особливо корисна для розріджених даних; менший кут (ближчий до нуля) вказує на вищу схожість між векторами.

2. Кореляція Пірсона (Pearson Correlation) - показує, коли послідовність рейтингів зростає або спадає разом. Вона розглядається як центрована версія косинусної схожості (наприклад, косинусна схожість, коли обидва вектори мають середнє значення, рівне нулю).

3. Евклідова відстань (Euclidean Distance Similarity) - використовує відстань між двома векторами для обчислення схожості (відстані) між користувачами.

1.3.3. Перехід до класифікаційної задачі та використання глибокого навчання

Таким чином, рекомендація наукових платформ традиційно досліджувалася за допомогою контентних та колаборативних

рекомендаційних методів. Проте, іншим перспективним підходом є трактування цієї проблеми як задачі класифікації тексту з наперед визначеним набором категорій.

У контексті даної роботи застосовуватиметься саме цей підхід. Текст визначається як інформація, надана користувачем про подання (назва, ключові слова тощо), а набір можливих вихідних класів складається з множини конференцій та журналів, отриманих із використаного набору даних.

Незважаючи на перспективність, підхід глибокого навчання до проблеми відповідності наукових платформ залишається малодослідженим. Найближча робота, яка використовує глибоке навчання для академічних рекомендацій, застосовує його для пропозиції платформ для проведення наукових зустрічей (meetup events).

Дане дослідження засвідчує, що існуючі академічні РС мають спільні недоліки: вони часто специфічні до конкретного набору даних, що ускладнює порівняння результатів між різними системами та вимагає розробки уніфікованих стратегій оцінювання.

Метою цього проекту є застосування глибокого навчання для рекомендації наукових платформ, використовуючи назву, ключові слова, CCS-концепти та анотацію рукописів конференцій/журналів.

Висновки до розділу

У першому розділі було проведено цілісний аналіз предметної області, що дозволив сформувати концептуальне бачення побудови рекомендаційної системи для відбору фахових наукових конференцій. Дослідження показало, що контентно-орієнтований підхід є найбільш придатним для визначення релевантності між науковими роботами дослідника та тематикою конференцій. Було з’ясовано, що сучасні рекомендаційні системи активно еволюціонують у напрямі глибокої семантичної обробки тексту, що суттєво

підвищує точність визначення відповідності. Аналіз літератури засвідчив, що переважна більшість існуючих рішень орієнтується на загальні метрики схожості, що не враховують індивідуальних характеристик профілю науковця. У дослідженні виявлено обмеження класичних моделей рекомендації, які часто не справляються зі складністю багатовимірних текстових даних. Було встановлено, що застосування методів глибокого навчання відкриває можливості для побудови більш гнучких і точних моделей класифікації.

РОЗДІЛ 2. РОЗРОБКА МЕТОДОЛОГІЇ ТА ПРОЕКТУВАННЯ АРХІТЕКТУРИ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ ВІДБОРУ ПЛАТФОРМИ ДЛЯ ФАХОВИХ НАУКОВИХ КОНФЕРЕНЦІЙ

2.1. Проектування рекомендаційної системи

Кінцевою метою проєкту є рекомендація відповідних наукових платформ користувачеві на основі вхідної інформації про подання, такої як назва, анотація тощо. Оскільки ми розглядаємо завдання рекомендації як проблему класифікації, основою системи є набір побудованих класифікаторів. За своєю суттю, система є контентно-орієнтованою рекомендаційною системою. У цьому розділі детально описується проектування цієї системи.

Враховуючи вхідні дані та наперед визначену множину класів, класифікатор прогнозує, до якого класу належать ці вхідні дані. Залежно від кількості класів, класифікатор може бути бінарним або мультикласовим. Ключова ідея нашої системи полягає в тому, щоб розглядати кожну конференцію чи журнал як окремий клас та розробляти класифікатори для кожного з них.

Коли інформація про подання надходить до системи, кожен із цих класифікаторів повертає ймовірність того, що подання може бути віднесено до його конкретної наукової платформи. Ці результати класифікації консолідуються та подаються користувачеві як рекомендація.

З міркувань продуктивності ми обрали бінарні класифікатори замість мультикласових. Ми розробили два підходи до проектування класифікаторів:

- Класифікатори окремих платформ (Single Venue Classifiers).
- Групові класифікатори (Group Classifiers).

У підході “класифікатори окремих платформ”, маючи загалом близько 309 наукових платформ, ми будуємо 309 унікальних бінарних 1D CNN класифікаторів (одномірних згорткових нейронних мереж), по одному для

кожної платформи. Такий вибір забезпечує репрезентацію кожної платформи в нашому корпусі.

Вхідний запис, для якого має бути надана рекомендація, передається до усіх 309 класифікаторів. Вони можуть працювати паралельно для забезпечення швидкого часу відгуку. Класифікатори, чиї платформи тісно пов'язані з вхідним записом, нададуть позитивні прогнози. Решта класифікаторів нададуть негативні прогнози.

Ці прогнози консолідуються та ранжуються відповідно до їхніх імовірностей і пропонуються кінцевому користувачеві. Крім того, для кожної рекомендованої платформи існує компонент схожості, який ідентифікує платформи, тісно пов'язані з прогнозованою.

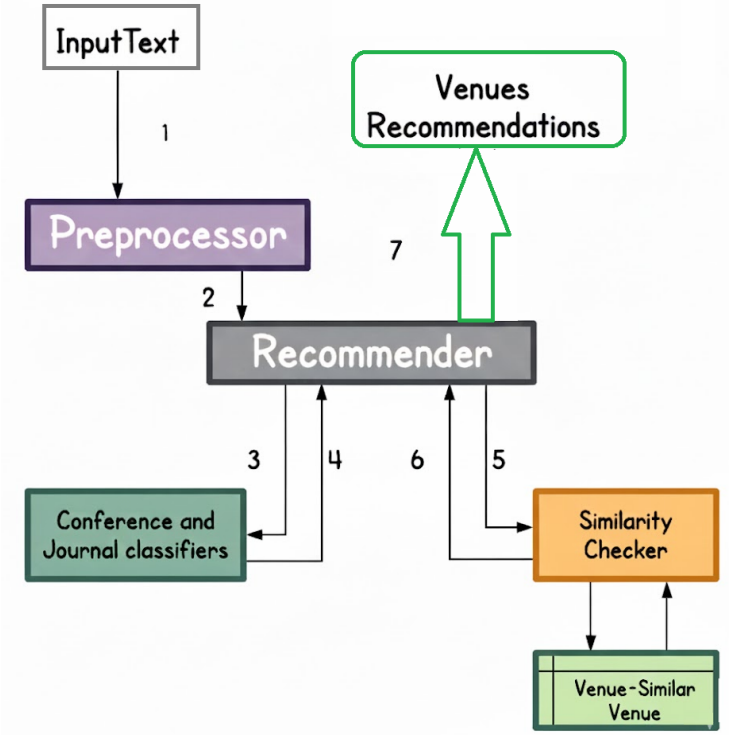

Рис. 2.1. Підхід класифікатора окремих платформ

Альтернативний підхід групових класифікаторів полягає в групуванні споріднених платформ і побудові одного бінарного 1D CNN класифікатора на групу. Ми ідентифікували загалом 56 груп. Для вхідного запису може бути рекомендована більше ніж одна група. Якщо група отримує позитивний

прогноз, усі платформи, що належать до цієї групи, рекомендуються користувачеві. Результати ранжуються відповідно до імовірностей класифікатора.

З точки зору користувацького досвіду, пропозиція групи може бути корисною, оскільки в межах групи існує суміш дедлайнів, місць проведення, розмірів конференцій та інших важливих атрибутів, які автор може розглянути. Іншою перевагою цього підходу є отримання кращої репрезентації схожих платформ, оскільки група сама по собі є їхньою колекцією. Отже, тут не потрібен окремий компонент схожості.

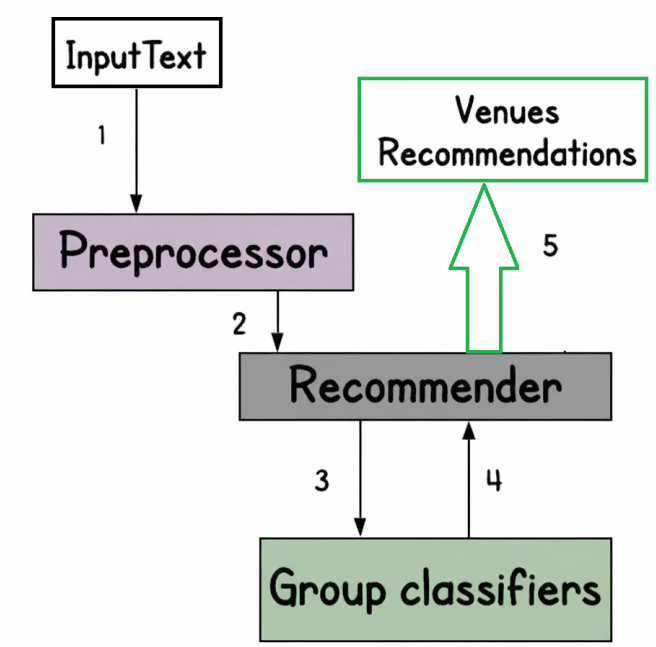

Рис. 3.2. Підхід групових класифікаторів

2.2. Вибір та обґрунтування моделі класифікації на основі глибокої нейтронної мережі

У цьому розділі обговорюється архітектура моделі, яка була використана для розробки класифікаторів у рамках даного дослідження.

Глибока нейронна мережа (DNN) є підкласом штучних нейронних мереж (ANN), що характеризується наявністю більш ніж одного прихованого шару між вхідним та вихідним шарами. Клас DNN охоплює широкий спектр

архітектур, включаючи згорткові нейронні мережі (CNN), рекурентні нейронні мережі довгої короткочасної пам'яті (LSTM), автокодувальники (Auto Encoders) та інші.

На високому рівні ANN є сукупністю штучних нейронів (вузлів). Різноманітність архітектур нейронних мереж базується на різних способах взаємозв'язку шарів цих одиниць.

Згорткові нейронні мережі (CNN) — це специфічний тип ANN, що містить один або кілька згорткових шарів (які не є повністю зв'язаними).

CNN здатні ефективно обробляти дані, що мають топологію, подібну до сітки [14]. Вони використовують операцію згортки (convolution operation) замість традиційного матричного множення принаймні в одному зі своїх шарів. Згортка є лінійною операцією над двома функціями дійсного аргументу. Спочатку CNN були успішно застосовані у сфері комп'ютерного зору, але згодом продемонстрували обнадійливі результати і в галузі обробки природної мови (NLP).

Ключовою перевагою CNN є їхня здатність виявляти просторові та часові ознаки у вхідному наборі даних за допомогою застосування фільтрів. В останні роки дослідження зосередилися на використанні CNN для класифікації тексту [26]. Наприклад, у роботі [15] класифікація речень здійснюється за допомогою CNN, які можуть ідентифікувати шаблони в тексті незалежно від їхнього позиційного розташування.

У даній дипломній роботі для розробки як одиночних, так і групових класифікаторів використовуються одномірні CNN (1D CNN). Одновимірна згорткова нейронна мережа (1D CNN) — це спеціалізований тип згорткової нейронної мережі (CNN), який використовується для аналізу послідовних або часових даних.

Головна відмінність 1D CNN від більш поширених 2D CNN (які використовуються для обробки зображень) полягає у вимірі, в якому виконується операція згортки.

Архітектура яка найкращим чином ілюструє, як CNN можуть бути застосовані для класифікації тексту показано на рис. 2.3.

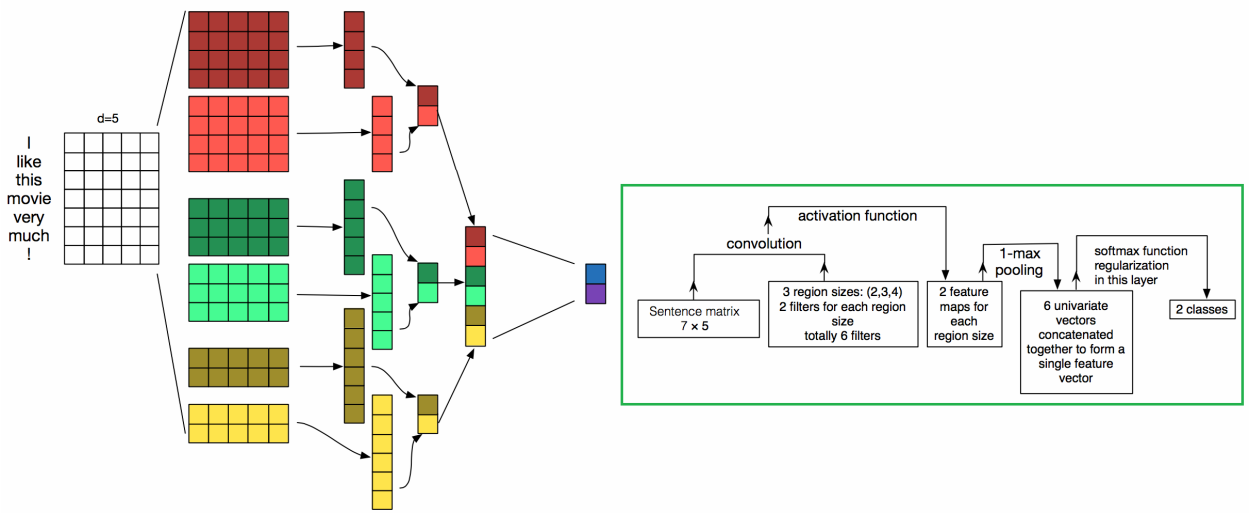

Рис. 2.3. Архітектура одномірної загорткової нейронної мережі

2.3. Розробка класифікаторів рекомендаційної системи на основі одновимірних нейронних мереж

У цьому розділі представлена детальна методологія розробки класифікаторів для кожної наукової платформи з використанням одновимірних згорткових нейронних мереж (1D CNN). Оскільки це прикладна система глибокого навчання, процес розробки включає послідовний конвеєр обробки даних, описаний нижче.

2.3.1. Консолідація та очищення даних

Наше дослідження базується на наборі даних бібліотеки ACM Digital Library (ACM-DL), представленому у форматі XML-файлів. Основне завдання цього етапу — перетворення сирих даних у структуру, придатну для подальшого моделювання. Детальний опис попередньої обробки даних буде надано дальше в роботі.

Дані ACM про конференції та журнали є специфічними для видання. Наприклад, файл SIGIR-2000 містить інформацію про всі статті, опубліковані

на конференції SIGIR у 2000 році. Оскільки інформація про серію конференцій або журнал може бути розподілена між багатьма XML-файлами (наприклад, SIGIR-14, SIGIR-13), першим кроком була консолідація всієї доступної інформації про конкретну серію конференцій або журнал в один уніфікований файл для забезпечення керованості даних.

На цьому етапі також відбувається вилучення релевантної інформації про подання. Створюється окремий файл для кожної наукової платформи, який являє собою колекцію пар:

<Вхідний_запис, Цільова_мітка>.

Цільова мітка: абревіатура наукової платформи.

Вхідний запис: конкатенація назви подання, анотації, ключових слів та набору концепцій CCS (Computing Classification System).

2.3.2. Попередня обробка даних

Після консолідації ми отримуємо дані для кожної платформи у форматі колекції записів

<Назва, Анотація, Ключові слова, CCS, Цільова мітка>.

Для навчання бінарного класифікатора дані представлені у двох класах:

- Позитивний клас (Цільова мітка = 1) - містить записи лише для конкретної платформи, яку моделює класифікатор.

- Негативний клас (Цільова мітка = 0) - містить записи всіх інших платформ.

Текстовий вхід для обох класів підлягає обрізанню для видалення стоп-слів та знаків пунктуації. Далі текст векторизується в набір цілих чисел за допомогою токенізатора. Внутрішньо токенізатор створює словник для

відображення текстових токенів на цілі числа. Об'єкт токенізатора зберігається для використання на етапі тестування.

2.3.3. Навчання моделі

Нейронні мережі оперують числовими даними, тому критично важливим етапом перед навчанням є векторизація тексту. Ми обрали векторизацію на рівні слів, що відповідає природі наших даних.

Для представлення слів використовувалися векторні вкладення (embeddings), а не one-hot кодування. Вкладення забезпечують компактне представлення слів як векторів, відображаючи семантичну інформацію в просторі: слова, близькі у векторному просторі, мають схоже значення. Вибір вкладень також виправданий короткою довжиною назв та кодів CCS.

Ми вирішили навчати вкладення безпосередньо з нашого корпусу, а не використовувати попередньо навчені моделі (Word2Vec, GloVe). Це рішення зумовлено тим, що слова в публікаціях ACM часто мають спеціалізоване значення, яке може відрізнятися від загального словника.

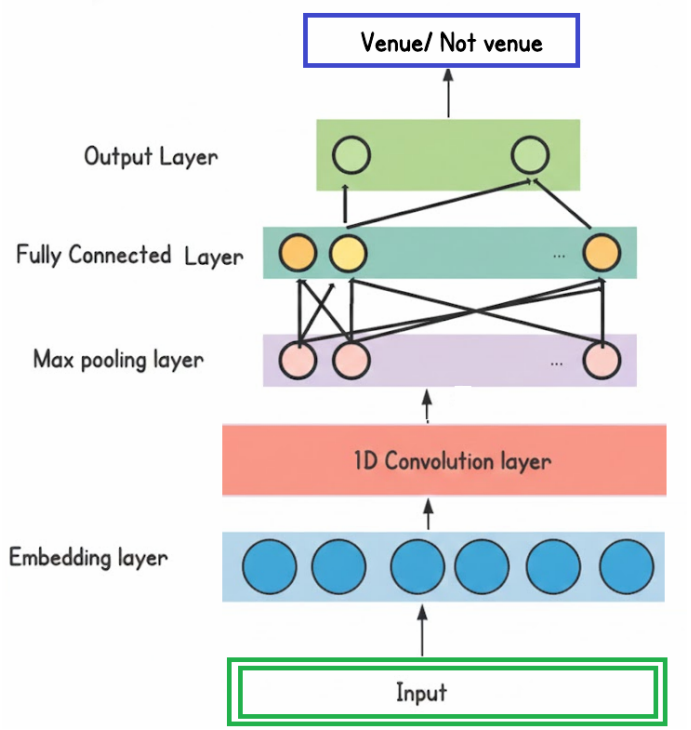

Рис. 2.4. Архітектура мережі для рекомендаційної системи

Для навчання моделей CNN була використана бібліотека глибокого навчання Keras. Архітектура CNN-моделі є спільною для одиночних та групових класифікаторів (як показано нарис. 2.4).

Кожен класифікатор навчався приблизно 30 епох для мінімізації перенавчання. Як функція втрат використовувалася бінарна перехресна ентропія у поєднанні з оптимізатором Adam.

Функція активації ReLU застосовувалася на шарі 1D CNN та щільному шарі. На вихідному шарі використовувалася сигмоїдна функція активації для отримання ймовірності класифікації.

2.3.4. Оцінка

Оцінювання класифікаторів здійснювалося за допомогою метрики F1-score. Детальні стратегії оцінки класифікаторів та рекомендаторів будуть викладені у третьому розділі.

Оскільки подібні рекомендаційні роботи з використанням даних бібліотеки ACM Digital Library не були ідентифіковані, у нас відсутня пряма базова лінія для порівняння. Порівняння з іншими дослідженнями вимагало б порівняння на різних наборах даних (CiteSeerX, IEEE, DBLP тощо). З огляду на це, оцінювання продуктивності системи ґрунтується на широкому наборі метрик, включаючи точність (accuracy), відгук (recall), F1-score та MAP (Mean Average Precision). Фактичні наукові платформи, на яких були опубліковані подання, використовуються як істинні мітки.

2.4. Характеристика та організація набору даних для дослідження

Для проведення дослідження були використані метадані про конференції та журнали, бібліотеки цифрових ресурсів асоціації обчислювальної техніки (ACM Digital Library, ACM-DL). Важливо, що надані метадані включають роботи, опубліковані також на конференціях, що не

належать до ACM. Дані представлені у форматі XML та містять файли конференцій та файли журналів.

Набір даних ACM структурно поділяється на дві основні категорії: матеріали конференцій (Conference Proceedings) та періодичні видання (Periodicals).

Кожен файл у корпусі матеріалів конференцій описує конкретне видання однієї конференції. Наприклад, файл ICML12 є XML-документом, що містить інформацію про конференцію ICML 2012 року. Інформацію в такому файлі можна логічно розділити на дві секції: 1) Секція метаданих та 2) Секція змісту.

Секція метаданих визначає властивості відповідної конференції. Нижче перелічено ключові поля (XML-теги), які є релевантними для рекомендаційної системи:

- conference rec
    - start date
    - end date
    -city
- proceeding rec
    - id
    - acronym
    - proc_desc
    - proc_title
    - proc_class
    - publication date
    - publisher
    - ID, code, name (для категорій)

Назва конференції у повному вигляді є досить розлогою і формується шляхом конкатенації даних із полів, наприклад як-от proc_title, proc_desc та proc_class.

Таблиця 2.1.

Приклади полів XM

Поле XML	Значення
proc_id	1116951
acronym	ILeGE-W'03
proc_desc	Матеріали 1-го міжнародного семінару LEGE-W
conference_number	1
proc_class	семінар
proc_title	Освітні моделі для послуг на основі GRID

З таблиці 2.1 видно, що повна назва конференції ("Матеріали 1-го міжнародного семінару LEGE-W з освітніх моделей для послуг на основі GRID") формується комбінуванням полів proc_desc та proc_title.

Для класифікаторів використовується:

- acronym (наприклад, "ILeGE-W") служить цільовою міткою (класом).

- Повні назви використовуються для ідентифікації подібних наукових платформ.

Секція змісту містить інформацію про статті, прийняті на конкретній конференції. Один файл конференції може містити від 10 до приблизно 50 статей. Кожне подання або стаття ідентифікується за допомогою тегу <article>.

Ключові поля у секції змісту для статті:

- article_id
- article publication date
- title
- ccs
- abstract
- keywords
- authors
- references

Більшість полів є самодокументованими. Винятково важливим є поле ccs (концепції CCS). Система обчислювальної класифікації ACM 2012 (CCS) широко застосовується для категоризації подань на конференціях ACM. Мета CCS — асоціювати кожне подання з релевантними категоріями, корисними для авторів, рецензентів та індексації в ACM-DL. Кожен термін CCS представляє концепцію в обчислювальній дисципліні. Система має шестирівневу поліієрархічну структуру , що відображено на рисунку 2.5.

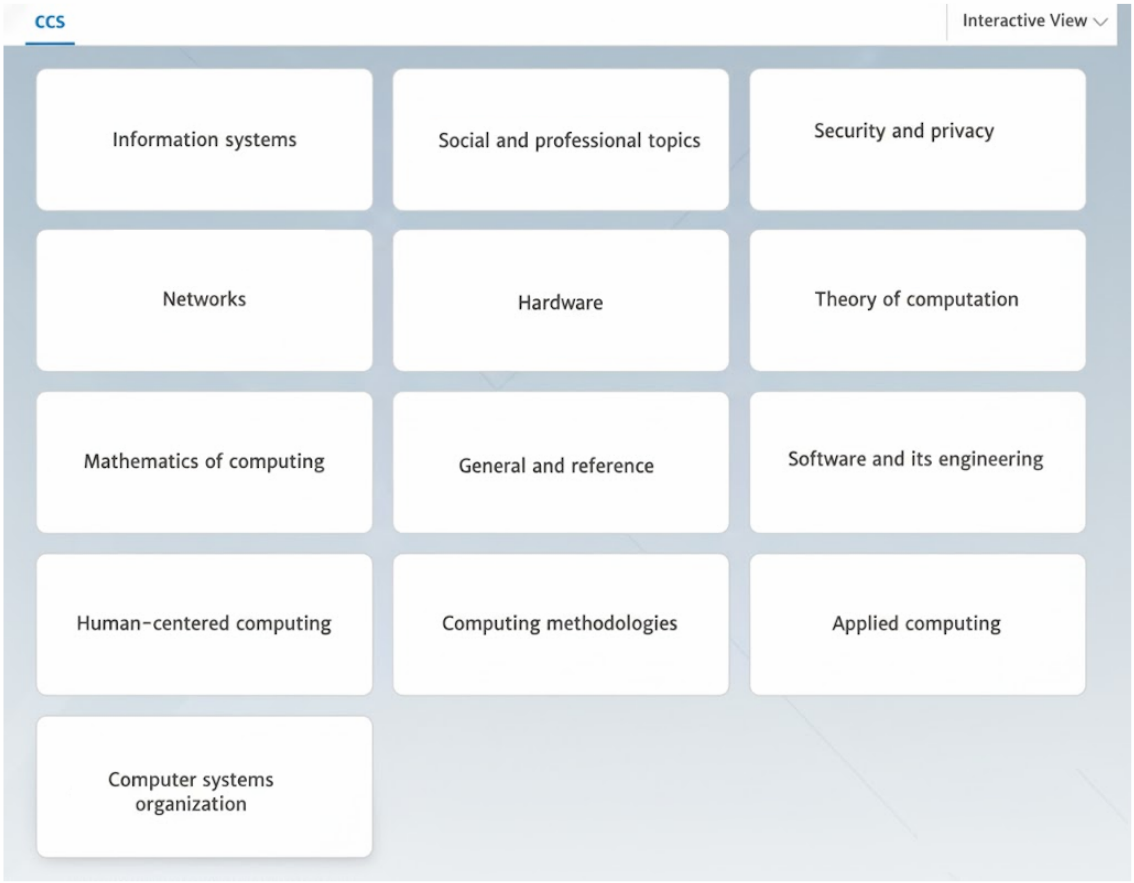

Рис. 2.5. Категорії в класифікаторі

Концепції CCS, визначені полем ccs статті, є критичним класом ознак для класифікації подань. Поля title, keywords, ccs та abstract із секції змісту формують сукупність вхідних ознак для завдання класифікації.

Друга категорія записів містить інформацію про журнали. Подібно до записів конференцій, кожен XML-файл журналу описує конкретне видання одного журналу та містить секції метаданих і змісту.

Важливі поля метаданих журналів:

- journal_id
- journal code
- journal name
- publisher
- id
- code
- city

Поле journal code (абсолютна абревіатура журналу) слугує вихідною міткою для класифікації журналів.

Секція змісту журналу містить статті з аналогічними ключовими полями, що й конференційні подання:

- article_id
- article publication date
- url
- title

Як і у випадку з конференціями, поля title, keywords, ccs та abstract використовуються як вхідні дані для навчання класифікаторів.

2.5. Етапи попередньої обробки даних та трансформація

Процес попередньої обробки даних (pre-processing) спрямований на конвертацію сирих даних у формат XML, описаних у попередньому розділі, у структуру, придатну для подальшого навчання моделей машинного та глибокого навчання. Ключовими етапами цього процесу є розбиття даних (data chunking) та очищення даних (data cleansing).

Набір даних мав значний обсяг. Для полегшення роботи та підвищення керованості дані були розбиті на кілька менших папок.

На етапі створення уніфікованого набору даних, метадані, розподілені в XML-файлах, трансформуються у формат CSV. Оскільки інформація про

одну наукову платформу (наприклад, серію конференцій) розподілена між кількома XML-файлами, здійснюється її консолідація в один CSV-файл для кожного формату платформи.

CSV-файл конференції містить наступні поля:

<Номер, Proc Id, Абревіатура, Рік, Назва матеріалів, Дата публікації, Видавець>

CSV-файл Журналу містить наступні поля:

<Номер, Код, Назва, Тип, Дата випуску>

Однією з головних методологічних проблем, ідентифікованих у наборі даних, була відсутність узгодженості у назвах наукових платформ.

Проблеми неузгодженості назв

- Змінність назви з часом. Назви конференцій можуть змінюватися. Наприклад, конференція "PAM" (Passive and Active Measurement) спочатку називалася "Пасивне та активне вимірювання", а потім її назву було змінено на "Пасивне та активне вимірювання мереж".

- Взаємозамінність назв. Для деяких конференцій (наприклад, "APNOMS") у наборі даних використовуються взаємозамінні повні назви.

- Включення надлишкової інформації. Назви часто включають специфічну інформацію про видання або назву видавця (наприклад, включення "ACM" або року конференції, попри наявність окремого поля "Рік" у метаданих).

Наявність незначних відмінностей у назвах платформ (які слугують цільовими мітками) є критичною перешкодою для побудови класифікаційної системи, оскільки це може призвести до неправильних припущень або помилкового навчання моделі. Відповідно, стандартизація назв платформ є необхідним кроком.

На етапі очищення даних здійснюється видалення надлишкових термінів із назв конференцій з метою виділення семантичної сутності (ядро) платформи. Був визначений набір так званих конференційних стоп-слів, що підлягають вилученню:

- Слова, що позначають тип публікації: "Матеріали", "Конференція", "Семінар".

- Слова, що описують подію: "Виклики".

- Порядкові числівники, що вказують на видання: "перший", "другий" тощо.

- Відомі назви видавців: IEEE, ACM.

- Інформація про рік (2024, 06, 6) також видаляється з поля назви.

Наприклад, повна назва "Матеріали міжнародної конференції 2025 року з моделювання та симуляції програмного забезпечення" після очищення перетворюється на "Моделювання та симуляція програмного забезпечення" (суть назви).

Таблиця 2.2.

Приклади відображення конференцій на семантичні терміни платформи

Абревіатура конференції	**Повна назва конференції**	**Терміни платформи (очищена сутність)**
ACL	Щорічна конференція Асоціації обчислювальної лінгвістики	Обчислювальна лінгвістика
ECCV	Європейська конференція з комп’ютерного зору	Комп’ютерний зір
SIGGRAPH	Група спеціальних інтересів з комп’ютерної графіки	Комп’ютерна графіка
SIGIR	Група спеціальних інтересів з пошуку інформації	Пошук інформації

Хоча очищені назви платформ відображають семантичну сутність, через ризик майбутніх змін у повних назвах, вони не були обрані як фінальні цільові мітки. Оскільки абревіатури конференцій/журналів залишалися постійними протягом років, було прийнято рішення використовувати їх як кінцеві цільові мітки для класифікаторів. Очищені назви платформ використовувалися лише для ідентифікації подібних наукових платформ.

2.6. Результати аналізу корпусу даних

Цей розділ присвячений обговоренню результатів, отриманих внаслідок дослідження структури та характеристик набору даних. Для полегшення аналізу, результати розділені відповідно до двох основних категорій: матеріали конференцій та періодичні видання.

Отриманий корпус даних містить інформацію про подання на конференції.

Розподіл подань за роками, проілюстрований на рисунку 2.6, демонструє поступове зростання кількості подань протягом зазначеного періоду. Пік кількості подань зафіксовано у 2010 році.

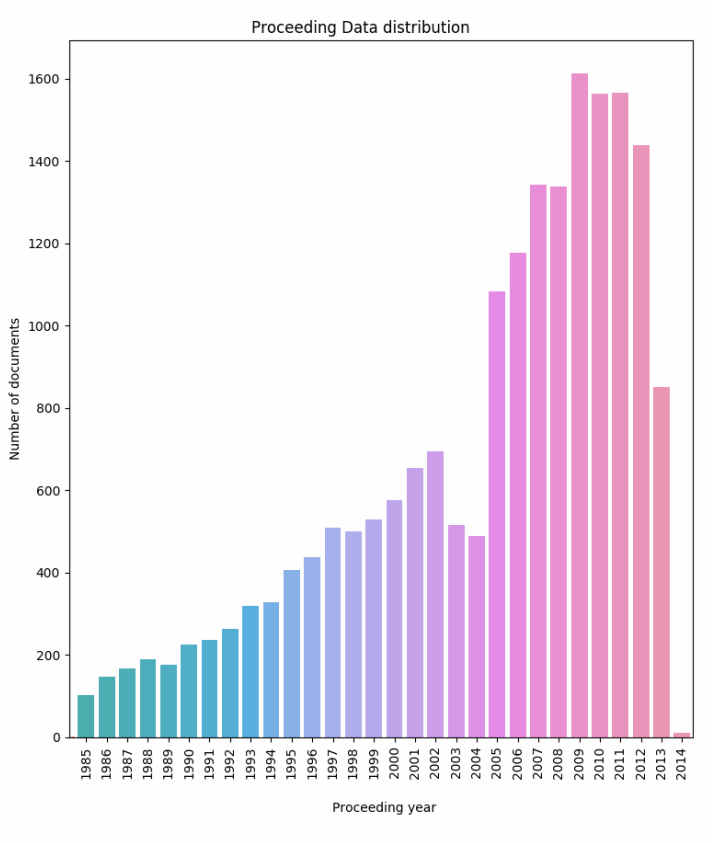

Рис. 2.6. Розподіл публікацій за роками в наборі даних

Із понад 4 тисяч конференцій майже половина конференцій не мали пов'язаної абревіатури. Оскільки ці випадки переважно стосуються старих

публікацій, вони були виключені з подальшого аналізу через їхню низьку релевантність для сучасної рекомендаційної системи.

Кожна конференція пов’язана з певним видавцем. Хоча дані отримані від ACM, корпус містить подання, опубліковані іншими організаціями, зокрема IEEE та Springer. Серед усіх конференцій, SIGGRAPH (спеціальна група інтересів з комп’ютерної графіки) має найбільшу кількість видань — 148. Ця кількість включає щорічні конференції, семінари, конкурси та супутні заходи. Наступною за кількістю йде HICSS (Гавайська міжнародна конференція з системних наук) зі 108 виданнями.

З метою забезпечення актуальності рекомендацій для поточного року, для навчання моделі будуть використовуватися лише дані, опубліковані в період з 2000 по 2014 рік.

У наборі даних ACM-DL було виявлено велику кількість файлів журналів. Кожен файл представляє конкретне видання певного журналу, тобто інформація про один журнал розподілена між кількома файлами. Загалом ідентифіковано 1300 унікальні журнали.

Серед унікальних журналів, TCS (Theoretical Computer Science) має найбільшу кількість видань — 823. Таблиця 2.3 містить інформацію про 10 журналів із найбільшою кількістю випусків.

Таблиця 2.3.

Журнали з найбільшою кількістю випусків

№ п/п	Журнал	Кількість випусків
1	Theoretical Computer Science	801
2	Communications of the ACM	575
3	Information Processing Letters	501
4	ACM SIGPLAN Notices	491
5	IEEE Transactions on Computers	470
6	Fuzzy Sets and Systems	461
7	The Computer Journal	454
8	Journal of Computational and Applied Mathematics	451
9	Fundamenta Informaticae	420
10	Journal of Computational Physics	401

Дані журналів також були класифіковані за типом видання.

Для цілей цієї магістерської роботи було встановлено критерії відбору наукових платформ:

- Наявність мінімум 200 подань у наборі даних.
- Платформа має бути актуальною та продовжувати публікуватися.
- Наявність пов'язаної абревіатури.

Застосування цих критеріїв призвело до формування фінального набору даних, що охоплює 309 наукових платформ (212 конференцій та 97 журналів). Гістограма розподілу даних у цих 309 платформах представлена на рис. 2.7.

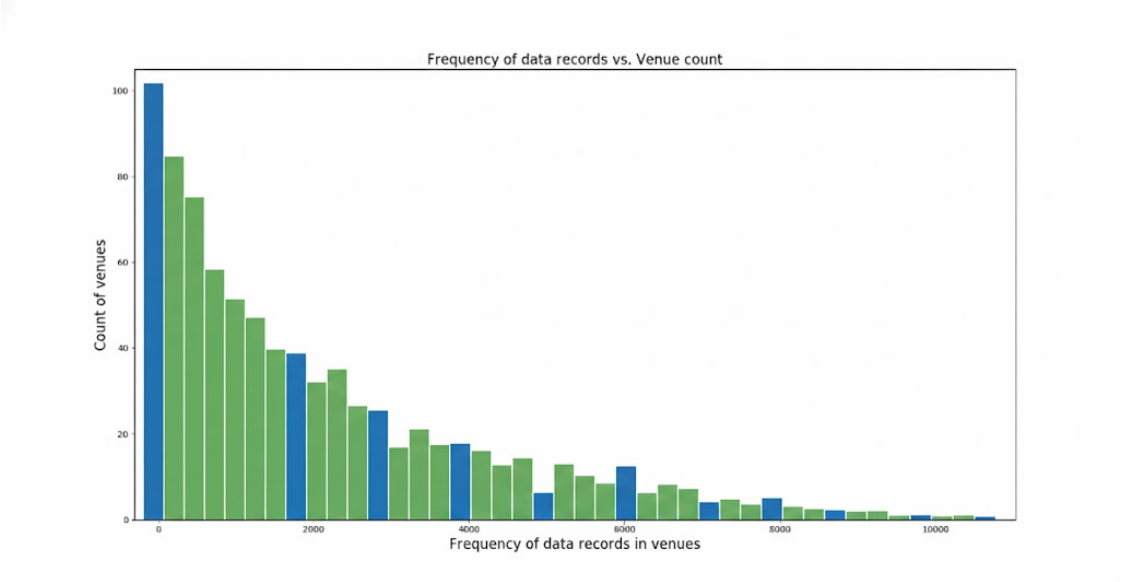

Рис. 2.7. Гістограма розподілу даних по платформах

Висновки до розділу

У другому розділі було сформовано методологію та спроєктовано архітектуру рекомендаційної системи, що ґрунтується на застосуванні глибоких нейронних мереж. Розробка моделі передбачала ретельний вибір типу нейронної мережі, в результаті чого зупинилися на одновимірній архітектурі як оптимальній для аналізу текстових даних. Методологічна частина дослідження охопила етапи підготовки даних, які виявилися

критичними для подальшого навчання моделі. Було продемонстровано, що консолідація, очищення та нормалізація текстів забезпечують стабільність і відтворюваність результатів класифікації. Значна увага була приділена розробці процедур трансформації даних, що дозволили подати текстові корпуси у вигляді структурованих числових представлень. Аналіз набору даних показав наявність суттєвого дисбалансу, що потребувало додаткових технік вирівнювання класів.

РОЗДІЛ 3. ОЦІНКА ТА ІМПЛЕМЕНТАЦІЯ МОДЕЛІ ДЛЯ ПОБУДОВИ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ ВІДБОРУ ФАХОВИХ НАУКОВИХ КОНФЕРЕНЦІЙ ВІДПОВІДНО ДО ПРОФІЛЮ НАУКОВЦЯ

3.1. Реалізація архітектури та прототипів рекомендаційної системи на основі підходу класифікаторів окремих платформ

Кожен бінарний класифікатор вимагає зразків даних для позитивного та негативного класів. Залежно від співвідношення між кількістю позитивних та негативних зразків, набір даних класифікатора може бути збалансованим або незбалансованим. Набір даних вважається збалансованим, коли кількість позитивних зразків дорівнює кількості негативних зразків. Ми експериментували з цим аспектом, розробивши три різні прототипи класифікаторів окремих платформ, оцінюючи їхню ефективність за допомогою показника F1-score.

Архітектура всіх прототипів CNN-класифікаторів є ідентичною, як показано на рисунку 3.1.

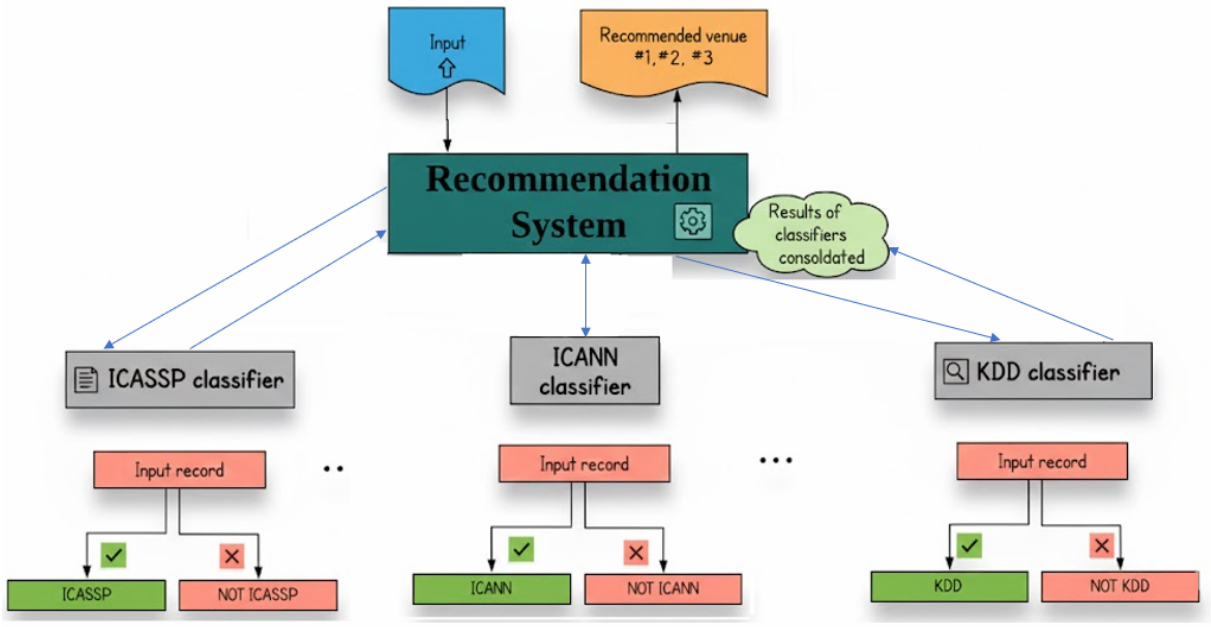

Рис. 3.1. Архітектура прототипів CNN-класифікаторів рекомендаційної системи

3.1.1. Представлення прототипу системи на основі базового незбалансованого набору даних

У прототипі P1 було розпочато побудову окремих бінарних класифікаторів для кожної платформи (веню). Розподіл даних у кожному класифікаторі був незбалансованим.

Позитивний клас: кількість зразків визначалася загальною кількістю даних для конкретної платформи у корпусі (наприклад, для KI — 605 записів).

Негативний клас: кількість зразків дорівнювала сукупній кількості записів усіх інших платформ.

Для класифікатора KI (605 позитивних записів), негативний клас містив близько 10 000 записів, що є класичним прикладом навчання на сильно незбалансованому наборі даних. Набір даних було розділено на навчальний, валідаційний та тестовий набори у співвідношенні 3:1:1. Вхідні ознаки включали конкатенацію назви подання, анотації, концепцій CCS та ключових слів, а вихідною міткою була абревіатура платформи. Цей підхід був протестований на вибірці з 30 класифікаторів.

Результати

Нерівномірна вибірка призвела до незадовільної продуктивності, а саме до дуже низьких показників F1-score (найвищий показник F1 становив 0.32). Через низьку точність класифікаторів подальше тестування рекомендаційної системи на основі P1 було визнано недоцільним. Було зроблено висновок, що використання набору даних із сильним переважанням негативного класу є неефективним.

3.1.2. Прототип системи на основі незбалансованого набору даних з обмеженням

Основною метою P2 було підвищення продуктивності класифікаторів. Ключова відмінність від P1 полягала у способі відбору зразків для

негативного класу. Було встановлено поріг на кількість записів, які можна вибрати з будь-якого негативного класу.

Для забезпечення мінімальної кількості даних для прогнозування, класифікатори будувалися лише для платформ, що мали понад 200 записів. Для формування негативного класу для будь-якого класифікатора вибиралося близько 200 записів з кожної іншої (негативної) платформи. Набір даних у P2 також залишався незбалансованим, але з меншим дисбалансом, ніж у P1.

На прикладі класифікатора KI (605 позитивних записів), негативний клас формувався шляхом вибірки близько 200 записів із кожної з решти платформ (встановлення порогу як 200 зразків на негативний клас).

Результати

P2 продемонстрував змішані результати щодо показників F1-score. Спостерігалося підвищення F1-score для деяких класифікаторів, проте проблема нестабільних та дуже низьких показників для інших класифікаторів збереглася. Було відзначено, що показники F1 для журналів були вищими, ніж для конференцій. Наприклад, показник F1 для позитивного класу класифікатора ICASSP становив 0.81, а для негативного класу — 0.98. Хоча P2 показав значне покращення порівняно з P1, проблема нестабільної продуктивності залишалася невирішеною.

3.1.3. Методологія побудови прототипу на основі збалансованого набору даних

Оскільки експерименти з незбалансованими наборами даних (P1, P2) не дали стійких результатів, у P3 було прийнято рішення збалансувати кількість позитивних та негативних зразків для кожного класифікатора.

Для будь-якого класифікатора кількість записів у позитивному класі визначала кількість зразків як у позитивному, так і в негативному класах.

На прикладі класифікатора KI (605 позитивних записів), 605 записів для негативного класу були випадковим чином відібрані з усіх інших платформ.

Результати

Р3 забезпечив значно кращу та стабільну продуктивність. Більшість класифікаторів досягли показника F1-score вище 0.70 для обох класів. Це підтвердило, що збалансування позитивних та негативних зразків є оптимальною стратегією для досягнення вищих та надійніших показників F1.

На основі результатів цих експериментів було встановлено, що збалансований набір даних (підхід Р3) є найкращим для навчання класифікаторів. Це рішення було поширене як на класифікатори окремих платформ, так і на групові класифікатори.

3.2. Методика розробки компонента визначення подібності платформ для наукових публікацій

Хоча наукова стаття або подання має лише одну істинну платформу (веню), потенційно воно могло б бути опубліковане в інших подібних платформах. Ми визначаємо платформи як подібні, якщо вони приймають роботи зі схожої обчислювальної дисципліни. Всі розроблені прототипи (Р1, Р2, Р3) функціонують згідно з архітектурою, показаною на рис. 3.1, де для будь-якого вхідного запису більше одного класифікатора може повернути позитивний прогноз.

Спочатку передбачалося, що ця властивість може бути використана для автоматичного визначення подібних платформ: якщо дві платформи повертають позитивний прогноз для одного й того самого тестового запису, вони вважаються подібними. Однак, аналіз рекомендацій, отриманих від прототипу Р3, показав, що подібні платформи не рекомендувалися у більшості випадків. Наприклад, для запису "Progressive Embedding", опублікованого у "SIGGRAPH", істинна платформа була рекомендована, але подібні платформи (такі як CGI, GI тощо) — ні. Оскільки рекомендаційна система залежала виключно від класифікаторів для пропозиції подібних платформ, виникла потреба у розробці незалежного компонента подібності.

Завданням компонента подібності є, ґрунтуючись на наборі позитивних прогнозів від класифікаторів, визначити платформи, подібні до тих класів, які отримали позитивні прогнози. Це питання стало основою для подального експерименту № 2.

Оскільки самі дані не містять явної інформації про тип прийнятих робіт, назва платформи була обрана як відправна точка для визначення подібних веню. Назва часто містить інформацію про обчислювальну дисципліну (наприклад, "Європейська конференція з комп’ютерного зору (ECCV)" вказує на роботи, пов’язані з комп’ютерним зором). Використовувалася гіпотеза, що подібні платформи мають подібні терміни у своїх назвах.

Ми експериментували з двома метриками текстової відстані/подібності: Відстань переміщення слів (Word Mover's Distance, WMD) та Косинусна подібність (Cosine Similarity, CoSim).

3.2.1. Застосування методу “відстань переміщення слів” (WMD)

WMD використовує векторні вкладення слів (word embeddings) для визначення подібності між двома текстовими документами. Зважаючи на коротку довжину тексту в назвах платформ, було виявлено, що цей підхід не є адекватним для нашого завдання. WMD є потужним інструментом, оскільки він враховує не лише наявність спільних слів, але й семантичну схожість між різними словами.

Наприклад, WMD між двома подібними конференціями з комп'ютерного зору, ACCV та ECCV, становила 1.79. Відстань між ACCV та CVPR (також комп'ютерний зір) становила 2.4. Водночас, відстань між неподібними конференціями ACCV та ACL становила 2.7. Відсутність чіткого шаблону (наприклад, для подібних ACL та CICLING відстань становила 2.3) ускладнювала визначення єдиного порогу подібності для всіх пар платформ. Через це WMD була відхилена.

3.2.2. Косинусна подібність (CoSim)

Іншою протестованою метрикою була косинусна подібність (CoSim), яка продемонструвала помірний успіх у визначенні подібних платформ, оскільки подібні веню часто мають спільні терміни у своїх назвах. Був емпірично встановлений поріг: дві платформи вважалися подібними, якщо їхня подібність перевищувала 85%.

Однак цей підхід мав недоліки: наприклад, він міг вважати платформи, пов'язані з "Наукою про дані" та "Передачею даних", подібними, хоча вони такими не є.

Через виявлені проблеми, результати CoSim були використані лише як основа для ручної перевірки та визначення подібних платформ. Інформація про природу подань була додатково верифікована через веб-сайти відповідних конференцій та журналів. На підставі цих даних була створена фінальна таблиця подібних платформ.

Приклади встановлених подібних платформ:

ACL → COLING, EACL, CICLING, NLDB, DOCENG (всі платформи з обчислювальної лінгвістики).

SIGCSE → ACE, ITICSE, CSEET.

TNET → SIGCOMM, ICNP, IEEENETW, ICN, LCN тощо.

Розроблено компонент перевірки подібності, який приймає на вхід задану платформу v, звертається до створеної таблиці подібних платформ та повертає список веню, подібних до v.

Цей компонент був доданий до архітектури прототипів, що призвело до фінальної архітектури класифікатора окремих платформ. Визначені подібні платформи також були використані для ручного створення груп для архітектури групового класифікатора (де група є колекцією подібних платформ).

Таким чином, експеримент №2 дозволив удосконалити процес ідентифікації подібних платформ та інтегрувати його в загальну архітектуру рекомендаційної системи.

3.3. Оцінка впливу формату вхідних ознак на продуктивність рекомендаційної системи

Ми провели третій набір експериментів, зосереджений на оптимізації вхідних даних для рекомендаційної системи. Рекомендатор може теоретично приймати будь-яку інформацію про подання (назву, повний текст, анотацію), але в цьому дослідженні ми оцінюємо його ефективність лише за двома основними формами вхідних даних:

1) Назва подання.

2) Назва та анотація подання.

Цей експеримент спрямований на визначення, яка з трьох форм вхідних даних анотації забезпечує кращий показник F1-score для категорії оцінки 2 групових класифікаторів:

а) Повна анотація.

б) Найчастіші слова (на основі частотного профілю) з анотації.

в) Іменники та прикметники (на основі профілю частин мови, POS) з анотації.

Категорія оцінки 2 для груп розглядає як істинну групу, так і подібні групи як релевантні рекомендації серед отриманого списку платформ.

Важливо зазначити, що назва подання надається рекомендатору без змін, оскільки вона містить критично важливу інформацію. Ми обрали категорію оцінки 2, оскільки вона враховує подібні групи, що дозволяє краще зрозуміти продуктивність рекомендатора щодо подібних платформ.

Для проведення експерименту було створено два типи профілів для кожної платформи та групи:

- Частотний профіль - складається з термінів із частотою більше ніж 20 у назвах та анотаціях, присутніх у навчальних записах.

- Профіль POS - створений шляхом вилучення іменників та прикметників із назв та анотацій навчальних записів.

Таблиця 3.1 табулює перші кілька термінів у цих профілях для деяких прикладних платформ.

Таблиця 3.1.

Частотний профіль та профіль частин мови (POS)

Платформа	Топ-10 слів у частотному профілі	Перші 10 слів у профілі POS
ACCV	moving, self, structure, important, information, kernel, resolution, facial, depth, super	paper, new, method, paper, questions, task, pose, method, dense
CEDN	high, explore, college, feedback, preservice, problem, integrating, potential, years, relationship	plethora, research, evidence, paper, results, aim, paper, picture, evaluation, study
Обчислювальна лінгвістика (група)	verb, formal, constraint, generative, application, multiple, clustering, hybrid, mt, event	glance, spatial, uses, preposition, language, scene, descriptions, knowledge, representation
CS Education (група)	interest, introduce, review, pedagogical, paradigm, colleges, framework, active, user, evaluating	last, years, nondeterminism, fundamental, concept, paper, approach, teach, paper, experience

Як вхідні дані рекомендаційній системі надається назва та анотація. Назва не фільтрується. Експериментуються три форми вхідних даних анотації:

1. Повна анотація: назва та повна анотація подання передаються без змін.

2. Анотація з частотними ознаками: перед подачею в класифікатор, рекомендатор перевіряє, чи присутній окремий токен в анотації у частотному профілі платформи. Якщо так, токен зберігається; інакше — відкидається. Таким чином виконується вибір ознак на анотації.

3. Анотація з POS-ознаками: аналогічний процес, але використовується профіль POS для фільтрації токенів анотації.

Згідно з таблицею 3.2, для перелічених груп рекомендатор досягнув кращих показників MAP (Mean Average Precision), коли вхідними даними слугували назва та ознаки POS з анотації.

Таблиця 3.2.

Показники MAP групових класифікаторів для різних форм ознак анотації

Група	MAP повної анотації	MAP ознак POS	MAP частотних ознак
Обчислювальна лінгвістика	0.71	0.60	0.27
Комп’ютерний зір та розпізнавання образів	0.73	0.74	0.51
Машинне навчання	0.58	0.59	0.20
Паралельні та розподілені системи	0.50	0.55	0.27

У випадках "Комп’ютерний зір та розпізнавання образів" та "Машинне навчання" показник MAP для ознак POS був найвищим (0.74 та 0.59 відповідно).

Форма "Частотних ознак" продемонструвала найнижчу продуктивність у всіх випадках.

Таким чином, у фінальній архітектурі, коли назва та анотація надаються як вхідні дані, рекомендатор фільтрує токени в анотації за допомогою профілю POS.

Проведені експерименти мали вирішальне значення для удосконалення продуктивності рекомендаційної системи. Вони дозволили досягти кращих показників F1 для класифікаторів та покращити рекомендації подібних платформ. Ці ідеї обґрунтовують остаточний вибір архітектури, представленої в другому розділі.

3.4. Результати оцінки одиночних класифікаторів

У цьому розділі представлено детальний аналіз результатів, отриманих для категорій оцінки одиночних та групових класифікаторів. Додатково, ми представляємо приклади рекомендацій, наданих системою для тестових записів, та проводимо аналіз причин можливих невдач.

Спочатку опишемо показники середньої точності (MAP) для одиночних класифікаторів платформ, а потім — показники для випадково обраної вибірки тестових записів.

Для зручності оцінки класифікатори були об'єднані у 15 партій. У таблиці 3.3 представлено продуктивність рекомендаційної системи. Партії з 1 по 10 містять записи конференцій, тоді як партії з 11 по 15 — записи журналів. Оцінка проводилася для двох форм вхідних даних:

- Тільки Назва (Назва подання).
- Назва + Анотація (Назва з функціями POS з анотації).

Таблиця 3.3.

Середня точність (MAP) для одиночних класифікаторів

Партія	**Кількість записів**	**Тільки назва: категорія 1**	**Тільки назва: категорія 2**	**Назва+анотація: категорія 1**	**Назва+анотація: категорія 2**
1	21	0.23	0.50	0.12	0.52
2	12	0.13	0.47	0.21	0.45
3	22	0.07	0.43	0.15	0.62
4	13	0.15	0.44	0.21	0.56
5	18	0.15	0.56	0.34	0.77
6	15	0.13	0.31	0.20	0.50
7	18	0.10	0.45	0.29	0.57
8	17	0.18	0.42	0.18	0.47
9	16	0.10	0.31	0.23	0.57
10	17	0.13	0.37	0.17	0.58
11	16	0.12	0.30	0.18	0.57
12	12	0.17	0.47	0.34	0.46
13	9	0.13	0.50	0.29	0.77
14	12	0.19	0.37	0.37	0.69
15	17	0.16	0.55	0.18	0.28

У таблиці 3.3 представлено порівняння середньої точності (MAP) одиночних класифікаторів для 15 партій тестових записів. Результати оцінюються за двома категоріями та двома формами вхідних даних:

Категорія 1 враховує лише істинну платформу як релевантну.

Категорія 2 враховує істинну та подібні платформи як релевантні.

Загалом, категорія 2 демонструє значно вищі показники MAP, підтверджуючи її корисність у рекомендаційній системі. Найкраща продуктивність (MAP 0.77) досягнута у партії 5 та партії 13 при використанні вхідних даних "назва + анотація" (де анотація фільтрується за ознаками POS).

Зважаючи на переваги категорії 2, для детального дослідження її продуктивності було проаналізовано показники F1@k для випадково обраних записів з усіх партій при різних значеннях k (таблиці 3.4 та 3.5).

Таблиця 3.4.

F1@k для одиночних класифікаторів для випадку “назва як вхідні дані”

ID назви	**k=1**	**k=10**	**k=20**	**k=30**	**k=40**
1	0.06	0.38	0.38	0.28	0.28
2	0.25	0.82	0.54	0.44	-
3	0	0.18	0.18	0.52	0.72
4	0.07	0.55	0.57	0.69	0.58
5	0	0.31	0.69	0.61	0.52
6	0.06	0.49	0.47	0.39	0.36
7	0	0.06	0.38	0.55	-
8	0.11	0.76	0.68	-	-
9	0.11	0.74	0.82	-	-
10	0.11	0.69	0.75	-	-
11	0.06	0.51	0.53	0.44	0.57
12	0.1	0.69	0.94	-	-
14	0.1	0.71	0.94	-	-
15	0.1	0.74	0.96	-	-
16	0.2	0.74	0.91	0.70	0.53
17	0	0.06	0.97	-	-
18	0	0.06	0.05	0.39	0.58
19	0	0.04	0.03	0.05	0.11
20	0	0.16	0.70	-	-

У таблиці 3.4 представлено показник F1@k для випадкової вибірки з 20 тестових записів, де вхідними даними класифікатора є лише Назва подання.

Порожні комірки ("-") вказують на відсутність отриманих рекомендацій при цьому значенні k. Спостерігається низька точність при k=1, оскільки більшість значень F1 є низькими або нульовими. Найкращі показники F1 (вище 0.90) досягнуті при k=20 для записів 12,13,14,15,16 та 17.

Низькі показники F1 при k=1 для обох форм вхідних даних пояснюються низькою точністю (Precision). Для обох форм вхідних даних значення k=20 демонструє кращі показники F1 порівняно з іншими значеннями k.

Таблиця 3.5.

F1@k для одиночних класифікаторів для випадку “назва та анотація як вхідні дані”

I ID назви	k=1	k=10	k=20	k=30	k=40
1	0.06	0.36	0.32	0.30	0.37
2	0.25	0.82	0.50	0.43	-
3	0.09	0.63	0.90	-	-
4	0.07	0.22	0.17	0.17	-
5	0	0.36	0.72	0.77	0.65
6	0	0	0.04	0.08	-
7	0	0	0.21	0.56	-
8	0.11	0.72	-	-	-
9	0.11	0.74	0.82	-	-
10	1.00	0.40	-	-	-
11	0.06	0.10	0.18	0.33	0.57
12	0.10	0.69	0.94	-	-
13	0.10	0.64	0.97	-	-
14	0.10	0.71	1.00	-	-
15	0.10	0.74	1.00	-	-
16	0.10	0.74	1.00	-	-
17	0	0.06	0.15	0.53	-
18	0.08	0.63	0.95	-	-
19	0	0	0.04	0.06	0.089
20	0	0.08	0.17	0.62	-

У таблиці 3.5 представлено показник F1@k для випадкової вибірки з 20 тестових записів, де вхідними даними класифікатора є назва та анотація (з

функціями POS). Порожні комірки ("-") вказують на відсутність рекомендацій. Порівняно з таблицею 3.4, при k=20 спостерігається ідеальна продуктивність (F1=1.00) для записів 14,15 та 16. Це підтверджує, що для більших k (зокрема k=20) використання відфільтрованої анотації разом із назвою може значно підвищити якість рекомендацій.

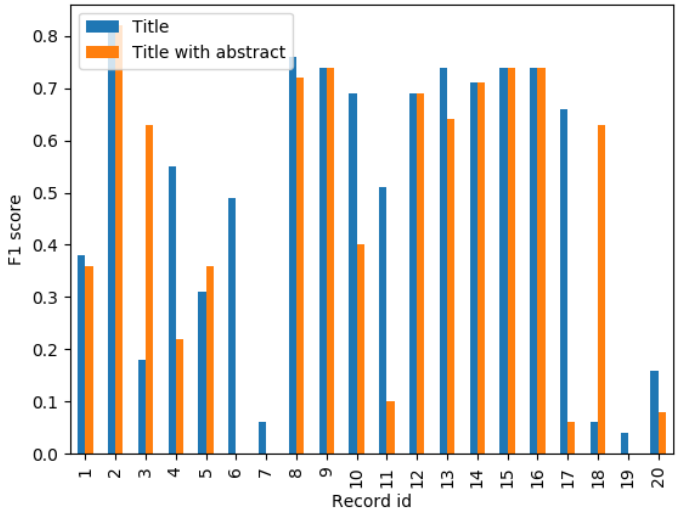

Рис. 3.2. Візуалізація F1-Score (F1@k=10)

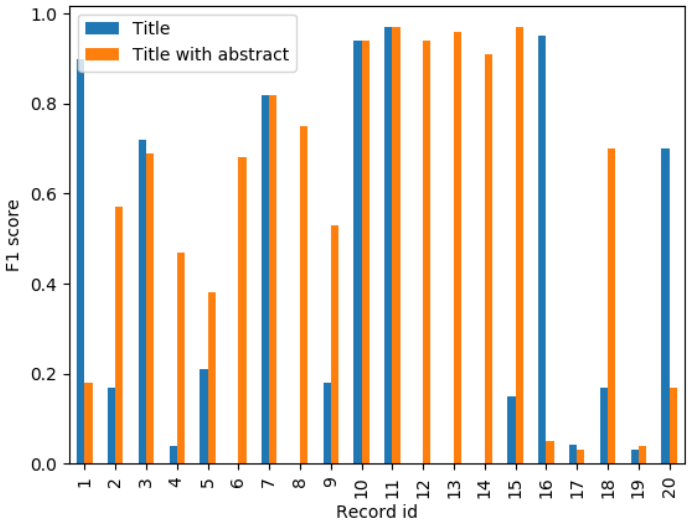

Рис. 3.3. Візуалізація F1-Score (F1@k=20)

Візуалізація F1-Score показано на рисунках 3.2 і 3.3 (F1@k=10) та (F1@k=20) відповідно.

Спостерігається, що для більшості записів перші 10 результатів (k=10) з тільки назвою як вхідними даними мають вищі показники F1. Ця тенденція змінюється для k=20, де використання назви та анотації як вхідних даних забезпечує кращі показники F1.

3.5. Результати продуктивності системи на основі підходу групових класифікаторів

У цьому розділі представлено оцінку продуктивності рекомендаційної системи на тестовому наборі для 56 груп платформ. Для тестування групових класифікаторів використовувалися ті самі тестові записи, що й для одиночних класифікаторів.

У таблиці 3.6 представлено найкращі середні показники точності (MAP) для двох основних архітектур (одиночні та групові класифікатори) відповідно до категорії оцінки 2 (включаючи істинні та подібні платформи/групи), що є найбільш практично значущим критерієм.

Таблиця 3.6.

Кінцеві результати продуктивності рекомендатора

Тип класифікатора	**Вхідні дані**	**Найкращий середній MAP (категорія 2)**	**Опис релевантності**
Одиночні (Single Classifiers)	назва + анотація	0.77	Істинна платформа та подібні платформи релевантні.
Групові (Group Classifiers)	тільки назва	0.42	Істинна група та подібні групи релевантні.
Групові (Group Classifiers)	назва + анотація	0.65	Істинна група та подібні групи релевантні.

Найкращий MAP (0.77) був досягнутий одиночними класифікаторами (Single Classifiers) у партії 5 та партії 13, коли використовувалися назва та анотація як вхідні дані, оцінюючись за категорією 2.

Як для одиночних, так і для групових класифікаторів, використання назви + анотації (з ознаками POS) призводить до значно вищих показників MAP у категорії 2, ніж використання лише Назви. Це підтверджує важливість фільтрації анотації за допомогою профілю POS.

Хоча одиночні класифікатори досягли найвищих пікових результатів (0.77), групові класифікатори також продемонстрували високий середній MAP (0.65) при використанні назви + анотації, що є важливим для надання рекомендацій щодо альтернативних місць публікації.

Партіоналізація (batches) в класифікаторах окремих платформ відрізняється від групування (groups) у групових класифікаторах. Таким чином, пряме порівняння між партіями одиночних класифікаторів та групами групових класифікаторів є некоректним.

Проте, обидві системи були протестовані на однаковому наборі тестових записів. Тому було обчислено середнє значення MAP для всіх партій і груп в обох типах класифікаторів.

Для коректного порівняння обрано наступні категорії оцінки:

- Категорія 5 (групи) vs. категорія 2 (одиночні): обидві ці категорії вважають істинну платформу та подібні платформи релевантними.

- Категорія 3 (групи) vs. категорія 1 (одиночні): обидві ці категорії вважають лише істинну платформу релевантною.

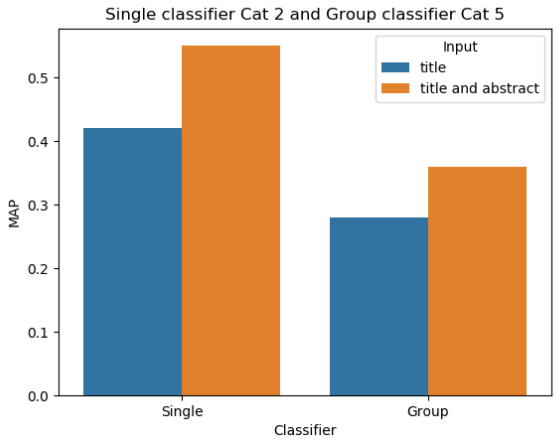

Рис. 3.4. Середній показник MAP для групової категорії 5 та одиночної категорії 2

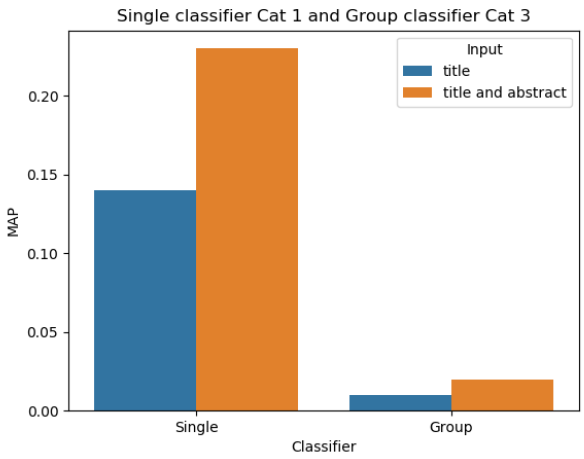

Рис. 3.5. Середній показник MAP для групової категорії 3 та одиночної категорії 1

Результати цього порівняння представлені на рис. 3.4 і 3.5.

При порівнянні зазначених вище категорій одиночний класифікатор демонструє кращу продуктивність.

Якщо ж порівнювати середній MAP категорії 2 для групових класифікаторів із середнім MAP категорії 2 для одиночних класифікаторів, то їхні значення є близькими: 0.51 та 0.55 відповідно.

Незважаючи на те, що одиночний класифікатор видається кращим варіантом з точки зору точності, його запуск для всіх партій є часозатратним процесом.

На нашому тестовому обладнанні надання рекомендацій займає в середньому близько 15 хвилин. Таке значне уповільнення є очікуваним зважаючи на велику кількість класифікаторів, які потрібно запустити. З точки зору практичного використання, групові класифікатори пропонують рекомендації значно швидше — приблизно за 4 хвилини.

Таким чином, вибір між архітектурами залежить від пріоритетів: одиночний класифікатор забезпечує дещо вищу точність, тоді як груповий класифікатор пропонує кращий компроміс між точністю та швидкістю у реальному часі.

У таблиці 3.7 наведено приклади топ-5 рекомендацій, наданих системою для всіх комбінацій вхідних даних. Для стислості таблиці, рекомендації подібних платформ/груп (з категорії 2) не включені.

Таблиця 3.7.

Приклади рекомендацій

ID запису	Одиночний класифікатор: тільки назва	Одиночний класифікатор: назва та анотація	Груповий класифікатор: тільки назва	Груповий класифікатор: назва та анотація
R1	COLING, **ACL**, ICNC, IJCAI, ICCS	COLING, IJCAI, ECCV, ITPM, Data	**Computational Linguistics**, Networks, HCI, Computational Science, Data	HCI, AI, Data, **Computational Linguistics**, ICCS
R2	**ECCV**, SIGGRAPH, FUNI, ICIC, SIGGRAPH	**ECCV**, ITPM, Image Processing, Robotics, Computer Graphics	**Computer Vision**, ML, Neuroscience, Robotics, Computer Graphics	**Computer Vision**, Image Processing, Multimedia, Robotics, HCI
R3	COMPSAC, **CRYPTO**, EUROPAR, MFCS, ICDT	**CRYPTO**, FOCS, COMPSAC, Robotics, OS	**Security & Privacy**, Networks, Math, Software Engineering, Medicine	**Security & Privacy**, HCI, Networks, Software Engineering, OS
R4	**ACCV**, SWAT, ESOP, IWANN, Document Engineering	**ACCV**, ECOP, IWANN, Computational Science, GIS	HCI, **Security & Privacy**, Combinatorics, Computational Science, Document Engineering	**Computer Vision**, Image Processing, Document Engineering, Multimedia, GIS
R5	**COCOON**, COLT, Math, Combinatorics, AI	**COCOON**, COLT, Operations Research, Math, AI	AI, **Theory of Computing**, Combinatorics, Operations Research,	ML, Neuroscience, Math, AI, Theory of Computing

Приклад (R2). Коли на вхід подається лише назва (колонка 2), одиночні класифікатори з позитивними прогнозами — це платформи "ECCV" та "SIGGRAPH". Компонент подібності в одиночних класифікаторах ідентифікує платформи, подібні до них, але вони в таблиці не відображаються. Аналогічно, для групових класифікаторів перелічено топ-5 рекомендованих груп без уточнення подібних груп.

Ключові спостереження наступні:

Істинні платформи (Ground Truth Venues) виділені жирним шрифтом. Наприклад, для запису "R1" істинна платформа — "ACL", а істинна група — "Computational Linguistics".

Нерекомендовані істинні значення. Спостерігається, що для деяких записів, як-от "R1", одиночні класифікатори, яким надано Назву та Анотацію як вхідні дані, не рекомендують істинну платформу "ACL".

Пропуск топ-5: для деяких записів, наприклад "R6", істинна платформа/група взагалі не потрапила до топ-5 рекомендацій.

Отже, створено систему, яка використовує технології глибокого навчання для надання рекомендацій відповідних місць публікації для академічних подань. Система здатна обробляти різні форми вхідних даних, включаючи назву, ключові слова, анотацію та концепції CCS. Представлено дві архітектури класифікаторів, що використовують одномірні згорткові нейронні мережі (1D CNN) для класифікації:

Одиночні класифікатори розроблені для прогнозування належності до кожного окремої платформи, групові класифікатори - для прогнозування належності до груп подібних платформ.

Проведено оцінювання розробленої системи із застосуванням стандартних метрик, таких як точність (Precision), відгук (Recall), F1-показник та середня точність (MAP). Для детального аналізу продуктивності було визначено та використано різні категорії оцінки. Виконано серію експериментів, спрямованих на підвищення продуктивності системи, включаючи балансування навчальних даних, оптимізацію вхідних даних та методику ідентифікації подібних веню. Ці експерименти сприяли удосконаленню загальної архітектури системи.

Під час виконання проєкту було виявлено низку методологічних та технічних обмежень:

- Початкові набори даних характеризувалися значною незбалансованістю класів, що негативно впливало на точність

класифікаторів. Ця проблема була частково вирішена шляхом застосування методів балансування даних.

- Задача кількісного визначення подібних веню виявилася нетривіальною через відсутність явних, маркованих даних про їхню схожість. Для вирішення цього використовувалися косинусна подібність та експертне (ручне) визначення.

- Дослідження базується виключно на наборі даних ACM Digital Library. Це обмежує можливість прямого порівняння отриманих результатів з іншими рекомендаційними системами, які використовують альтернативні джерела даних.

- Наявність обмежених обчислювальних ресурсів вплинула на загальну кількість та обсяг експериментів, які могли бути проведені для подальшої оптимізації системи.

3.6. Опис пропонованої методології щодо побудови рекомендаційної системи вибору платформи фахових публікацій

Основною задачею яка вирішується це допомогти досліднику знайти ідеальне місце для публікації наукової роботи. Дана система рекомендацій використовує підхід глибокого навчання, що поєднує точність окремих рішень та широту групових рекомендацій.

1. Вхід: подання текстових даних

Все починається з академічного подання. Система приймає текст, що описує роботу:

- Назва (Title)

- Анотація (Abstract)

- Ключові слова (Keywords)

- Додаткові класифікатори, такі як концепції CCS (Computing Classification System).

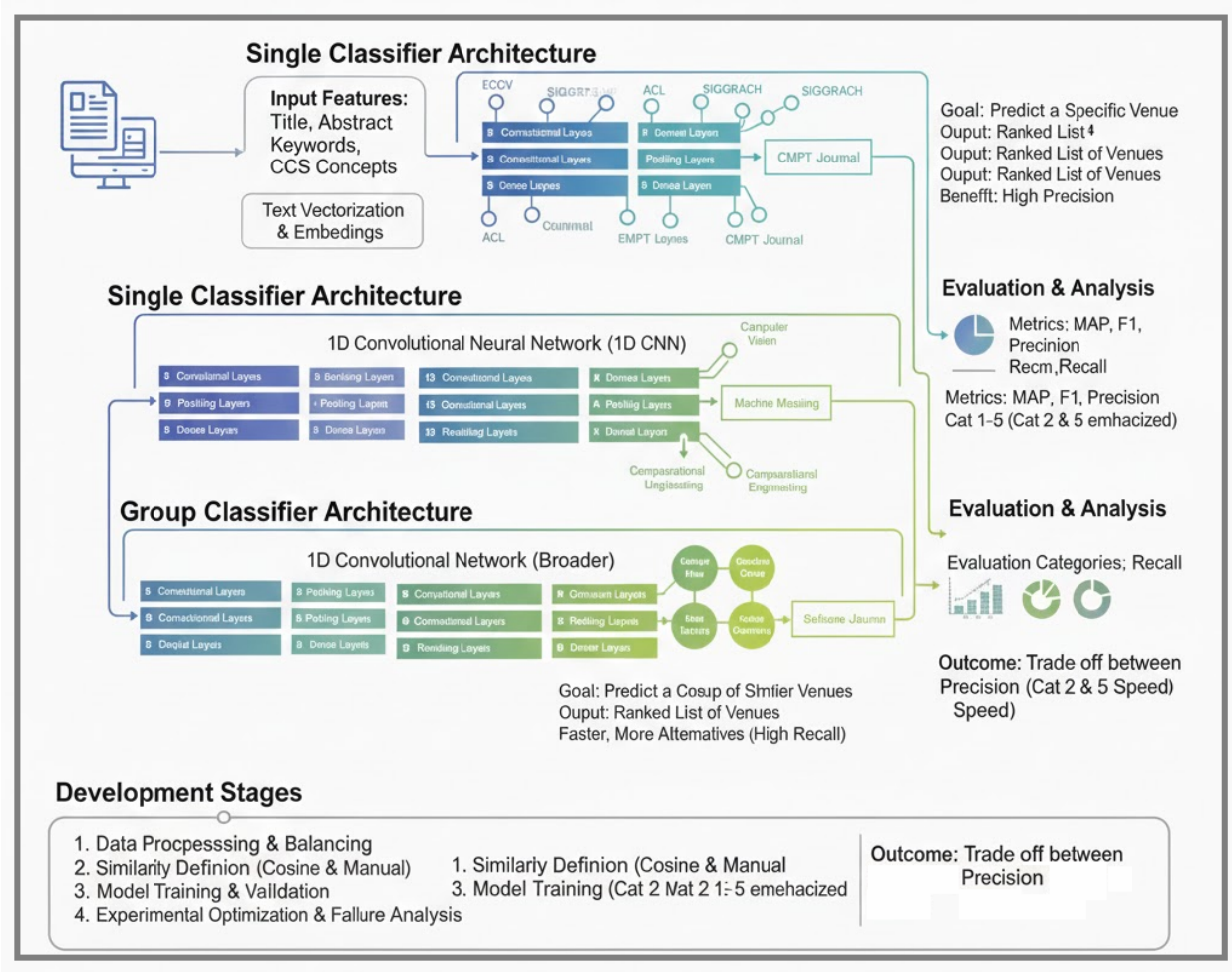

Рис. 3.6. Ілюстрація пропонованої методології побудови рекомендаційної системи

Цей сирий текст перетворюється на числовий формат (векторизація), готовий до обробки нейронними мережами.

2. Подвійний шлях класифікації

Наша методологія є двоканальною, але обидва канали використовують подібний механізм: одномірну згорткову нейронну мережу (1D CNN). Ця мережа спеціально розроблена для виявлення важливих, ієрархічних патернів у послідовностях, що робить її ідеальною для аналізу тексту.

А. Одиночні класифікатори

Це найбільш детальний і трудомісткий рівень. Замість однієї великої моделі, тут існує безліч бінарних класифікаторів, по одному для кожної потенційної платформи (веню) у наборі даних.

Кожен класифікатор вирішує просте питання: "Чи підходить ця робота для конференції X?"

Цей підхід забезпечує високу точність (Precision) прогнозу на рівні конкретної платформи.

Б. Групові класифікатори

Паралельно працює канал, який об'єднує подібні платформи у тематичні групи (наприклад, "Машинне Навчання", "Комп'ютерний Зір").

Цей класифікатор прогнозує найбільш релевантну тематичну групу для поданої роботи. Такий підхід значно швидший (4 хвилини проти 15 хвилин), оскільки моделюється менша кількість класів. Він ідеально підходить для надання альтернативних та схожих місць публікації, забезпечуючи кращий відгук (Recall).

3. Формування фінальної рекомендації

Система консолідує результати від обох архітектур. Навіть якщо одиночний класифікатор не включив істинну платформу до топ-5, він може підвищити її рейтинг завдяки компоненту подібності (косинусна подібність), який враховує схожі платформи.

4. Оцінка: баланс між точністю та швидкістю

Кінцева оцінка системи підтвердила важливий компроміс:

- Одиночні класифікатори досягають дещо вищої точності (MAP ≈0.55), але потребують багато часу.

- Групові класифікатори пропонують майже таку ж точність (MAP ≈0.51), але зі значно кращою швидкістю, що робить їх більш практичними для використання в реальному часі.

Висновки до розділу

У третьому розділі було здійснено експериментальну реалізацію запропонованої архітектури та проведено комплексну оцінку продуктивності

моделі. Було створено кілька прототипів рекомендаційної системи, що відрізнялися складом і збалансованістю наборів даних, що дозволило глибше зрозуміти вплив структури вибірки на результати. Експериментальні перевірки засвідчили, що одновимірні нейронні мережі демонструють високу точність при обробці коротких наукових анотацій і тематичних описів. Дослідження формату вхідних ознак підтвердило, що використання векторних подань значно покращує здатність моделі до семантичного узагальнення. Результати експериментів продемонстрували, що ансамблеві класифікатори здатні підвищувати стабільність моделі та покращувати узагальнювальну здатність системи. Реалізована методологія показала, що комплексний підхід до оцінки як класифікації, так і подібності дозволяє створити багатокомпонентну рекомендаційну систему з високою точністю.

ВИСНОВКИ

У ході виконання даної магістерської роботи було здійснено дослідження моделей, методів та технологічних засобів побудови рекомендаційної системи для автоматизованого відбору фахових наукових конференцій відповідно до профілю науковця. Результати трирівневого аналізу – теоретичного, методологічного та експериментально-практичного – дозволили сформувати науково обґрунтовану концепцію побудови рекомендаційних систем нового покоління, орієнтованих на потреби сучасних дослідників.

У першому розділі було проведено систематизований огляд сучасних рекомендаційних підходів та специфіки їх застосування у сфері наукових комунікацій. Деталізовано контентно-орієнтовані моделі, що лежать в основі рекомендаційних систем для вибору наукових публікаційних платформ, та визначено ключові характеристики, що формують профіль науковця й описують релевантність конференції.

Огляд літератури показав, що рекомендаційні системи у науковому середовищі зазвичай орієнтуються на текстовий контент анотацій, тематичну близькість та метадані публікаційних платформ, проте значна частина існуючих розробок не забезпечує індивідуалізованого відбору конференцій на основі персонального наукового доробку. Було встановлено, що класичні моделей рекомендації (контентні, колаборативні, гібридні) поступово доповнюються методами глибокого навчання, що істотно підвищують точність класифікації наукових профілів і платформ. Виявлено, що формальна постановка задачі рекомендації конференцій природно переходить у задачу багатокласової класифікації з використанням глибинних нейронних мереж, які здатні ефективно працювати з текстовими ознаками та латентними семантичними представленнями.

У другому розділі було сформовано методологію побудови рекомендаційної системи та спроєктовано її архітектуру. Обґрунтовано вибір

глибокої нейронної мережі як базової моделі класифікації через її здатність автоматично вилучати семантично значущі ознаки з текстових даних, включно з анотаціями статей та описами конференцій. Проведено порівняння можливих підходів до побудови класифікатора та вибрано архітектуру одновимірної нейронної мережі (1D-CNN), яка забезпечує оптимальне співвідношення точності, швидкодії та стійкості до надлишковості текстових ознак.

Особлива увага була приділена організації набору даних. Було здійснено багаторівневу очистку корпусу текстів, нормалізацію, токенізацію та трансформацію даних у формат, придатний для обробки нейронною мережею. Результати аналізу корпусу даних довели необхідність балансування вибірки та застосування спеціалізованих технік попередньої обробки, що мінімізують вплив шумових або нерепрезентативних фрагментів тексту.

У третьому розділі було здійснено імплементацію запропонованої архітектури та проведено експериментальну перевірку ефективності різних конфігурацій системи. Реалізовано прототипи на основі трьох варіантів даних: базового незбалансованого набору, незбалансованого набору з обмеженнями та збалансованого набору. Кожен із прототипів продемонстрував різні показники точності, повноти та F1-міри, що дозволило визначити оптимальні умови навчання моделі.

Оцінка впливу формату вхідних ознак показала, що результати класифікації помітно залежать від якості попередньої обробки та вибору текстових репрезентацій. Зокрема, використання ембеддингів дозволило моделі краще захоплювати латентні семантичні зв’язки.

У роботі також розроблено компонент визначення подібності між науковими платформами. Порівняння двох підходів — відстані переміщення слів (WMD) та косинусної подібності — продемонструвало, що WMD забезпечує глибшу семантичну оцінку, тоді як CoSim є швидшим і

практичнішим для великих даних. Поєднання цих методів дозволяє гнучко адаптувати систему до різних сценаріїв використання.

Здійснено детальну оцінку продуктивності одиночних класифікаторів та групових класифікаторів (ensemble-based). Підхід групових класифікаторів показав вищу стабільність та кращу узагальнювальну здатність, особливо при роботі з гетерогенними наборами даних. Це дало змогу створити методологію побудови рекомендаційної системи, яка включає симбіоз класифікаційних моделей та модулів семантичної подібності.

Отримані результати демонструють наукову новизну та практичну цінність запропонованої методології, оскільки створений підхід може бути використаний як основа для побудови повноцінних інтелектуальних систем підтримки вибору публікаційних платформ, а також адаптований для задач рекомендації журналів, наукових колаборацій чи тематичних дослідницьких подій.

ПЕРЕЛІК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Recommendation in Academia: A Joint Multi-Relational Model / Zaihan Yang // https://www.cse.lehigh.edu/~brian/pubs/2014/ASONAM/recommendation-in-academia.pdf
2. Publication Venue Recommendation Using Author Network's Publication History - https://www.researchgate.net/publication/262290496_Publication_Venue_Recommendation_Using_Author_Network%27s_Publication_History
3. Academic Venue Recommendations Based on Similarity Learning of an Extended Nearby Citation Network - https://www.researchgate.net/publication/331884701_Academic_Venue_Recommendations_Based_on_Similarity_Learning_of_an_Extended_Nearby_Citation_Network
4. Incorporating Full Text and Bibliographic Features to Improve Scholarly Journal Recommendation | IEEE Conference Publication | IEEE Xplore - https://ieeexplore.ieee.org/document/8791200
5. Recommendation of scholarly venues based on dynamic user interests – ScienceDirect - https://www.sciencedirect.com/science/article/abs/pii/S1751157716303406
6. Recommendation of Scholarly Venues Based on Dynamic User Interests - https://www.researchgate.net/publication/311736665_Recommendation_of_Scholarly_Venues_Based_on_Dynamic_User_Interests
7. Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." Journal of Machine Learning Research, vol. 3, 2003, pp. 993–1022.
8. Salakhutdinov, Ruslan, Andriy Mnih, and Geoffrey Hinton. "Restricted Boltzmann Machines for Collaborative Filtering." Proceedings of the 24th International Conference on Machine Learning (ICML), ACM Press, 2007, pp. 791–798.

9. Rendle, Steffen. “Factorization Machines.” 2010 IEEE International Conference on Data Mining, IEEE, Sydney, 2010, pp. 995–1000.
10. Liu, Bing. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, 2012.
11. McCallum, Andrew, et al. “The Author–Topic Model for Authors and Documents.” Proceedings of the 20th International Conference on Machine Learning, AAAI Press, Washington, DC, 2004, pp. 105–112.
12. Koren, Yehuda, Robert Bell, and Chris Volinsky. “Matrix Factorization Techniques for Recommender Systems.” Computer, vol. 42, no. 8, IEEE, 2009, pp. 30–37.
13. Mikolov, Tomas, et al. “Efficient Estimation of Word Representations in Vector Space.” Proceedings of Workshop at ICLR, 2013.
14. Devlin, Jacob, et al. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” Proceedings of NAACL-HLT, ACL, 2019, pp. 4171–4186.
15. Vaswani, Ashish, et al. “Attention Is All You Need.” Advances in Neural Information Processing Systems (NeurIPS), 2017.
16. Joachims, Thorsten. “Text Categorization with Support Vector Machines: Learning with Many Relevant Features.” ECML 1998, Springer, Berlin, 1998, pp. 137–142.
17. Ricci, Francesco, Lior Rokach, and Bracha Shapira. Recommender Systems Handbook. Springer, 2015.
18. Hofmann, Thomas. “Probabilistic Latent Semantic Indexing.” Proceedings of the 22nd Annual International ACM SIGIR Conference, ACM, Berkeley, 1999, pp. 50–57.
19. Zhang, Shuai, et al. “Deep Learning Based Recommender System: A Survey.” ACM Computing Surveys, vol. 52, no. 1, ACM Press, 2019, pp. 1–38.
20. Yu, Keming, et al. “Large-Scale Bayesian Logistic Regression for Text Categorization.” Technometrics, vol. 48, no. 4, 2006, pp. 542–553.

21. Sugiyama, Kazunari, and Min-Yen Kan. “A Comprehensive Evaluation of Scholarly Paper Recommendation.” JCDL 2010, ACM, Gold Coast, 2010, pp. 305–314.
22. Beel, Joeran, Bela Gipp, and Erik Wilde. “Research-Paper Recommender Systems: A Literature Survey.” International Journal on Digital Libraries, vol. 17, Springer, 2016, pp. 305–338.
23. Huang, Sheng, et al. “Recommending Scholarly Venues Based on User Publication Records.” JCDL 2014, IEEE, London, 2014, pp. 147–156.
24. Chandramouli, Harsha V., et al. “A Content-Based Scholarly Paper Recommendation System.” IEEE International Conference on Big Data, IEEE, 2015, pp. 2413–2418.
25. Ren, Xiang, et al. “ClusCite: Effective Citation Recommendation by Information Network Embedding.” KDD 2014, ACM, New York, 2014, pp. 821–830.
26. Wang, Ding, et al. “A Survey on Scholarly Data: Citation Analysis, Topic Modeling, and Publication Recommendation.” IEEE Transactions on Big Data, vol. 7, no. 3, 2021, pp. 527–545.
27. Hamilton, William L., Jure Leskovec, and Dan Jurafsky. “Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change.” ACL 2016, Berlin, ACL, 2016, pp. 1489–1501.
28. He, Xiangnan, et al. “Neural Collaborative Filtering.” WWW 2017, ACM, Perth, 2017, pp. 173–182.
29. Khan, Shafiq R., et al. “A Survey of Scholarly Recommender Systems.” Scientometrics, vol. 125, Springer, 2020, pp. 441–469.
30. Ning, Xia, and George Karypis. “SLIM: Sparse Linear Methods for Top-N Recommender Systems.” 2011 IEEE International Conference on Data Mining, IEEE, Vancouver, 2011, pp. 497–506.
31. Khabsa, Madian, and C. Lee Giles. “The Number of Scholarly Documents on the Public Web.” PLOS ONE, vol. 9, no. 5, 2014, pp. 1–6.

32. Tang, Jie, et al. “ArnetMiner: Extraction and Mining of Academic Social Networks.” KDD 2008, ACM, Las Vegas, 2008, pp. 990–998.
33. Schein, Andrew I., et al. “Methods and Metrics for Cold-Start Recommendations.” SIGIR 2002, ACM, Tampere, 2002, pp. 253–260.
34. Beel, Joeran, and Bela Gipp. “Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar.” Journal of Scholarly Publishing, vol. 43, no. 2, 2012, pp. 176–190.
35. Bar-Ilan, Judit. “Which h-Index?—A Comparison of WoS, Scopus and Google Scholar.” Scientometrics, vol. 74, Springer, 2008, pp. 257–271.
36. Chen, Xin, et al. “Personalized Academic Paper Recommendation Using Citation Proximity and Co-authorship Information.” JCDL 2013, ACM, Indianapolis, 2013, pp. 313–322.
37. Wang, Peng, et al. “Learning Hierarchical Representation Model for NextBasket Recommendation.” SIGIR 2015, ACM, Santiago, 2015, pp. 403–412.
38. Wu, Yequan, et al. “A Neural Probabilistic Model for Document Representation.” EMNLP 2015, ACL, Lisbon, 2015, pp. 1369–1374.
39. Chen, Zhiyong, and Jun Yan. “Topic Modeling for Scholarly Recommendation and Retrieval.” Journal of Informetrics, vol. 7, Elsevier, 2013, pp. 798–810.
40. Qiu, Minghui, et al. “A Survey of Text Representation Methods for Deep Learning.” IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 10, 2022, pp. 4912–4932.
41. Wang, Xiaoyan, et al. “A Survey on Academic Venue Recommendation Systems.” Information Processing & Management, vol. 59, no. 3, Elsevier, 2022, pp. 1–18.