МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

Кваліфікаційна наукова
праця на правах рукопису

ХУДІК БОГДАН ОЛЕКСАНДРОВИЧ

УДК 004.9: 004.8: 004.65: 004.22

ДИСЕРТАЦІЯ

МЕТОДИКА ПОБУДОВИ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ ДЛЯ СФЕРИ ОСВІТИ НА ОСНОВІ ГІБРИДНОГО ПІДХОДУ

Спеціальність 123 «Комп'ютерна інженерія»
Галузь знань 12 «Інформаційні технології»
Подається на здобуття наукового ступеня доктора філософії

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело
_____Б.О.Худік
(підпис)

Науковий керівник Золотухіна Оксана Анатоліївна, кандидат технічних наук, доцент

Електронний підпис
ХУДІК
БОГДАН
ОЛЕКСАНДРОВИЧ
Ідентифікаційний
код
3437612915
Підписано у Вчасно

Київ 2023

АНОТАЦІЯ

Худік Б.О. Методика побудови рекомендаційної системи для сфери освіти на основі гібридного підходу. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 123 Комп'ютерна інженерія. – Державний університет інформаційно-комунікаційних технологій Міністерства освіти і науки України. – Київ, 2023.

Рекомендаційні системи представляють собою підклас систем прийняття рішень, які використовують знання про інтереси та переваги людини для оцінки та прогнозування її реакції на рекомендацію купити певний товар або скористатися деякою послугою. Сучасні підходи до організації освітніх процесів при формуванні індивідуальної освітньої траєкторії здобувачів освіти потребують формування певних рекомендацій щодо вибору навчальних курсів, навчальної літератури, прогнозування запису користувачів на той чи інший курс та вирішення інших допоміжних освітніх задач, що робить ці процеси сферою інтересу рекомендаційних систем. Аналіз сфери використання рекомендаційних систем показав, що більшість досліджень та розробок концентруються на забезпеченні комерційних задач та цілей, які пов'язані збільшенням доходів, розширенням клієнтської бази тощо, тоді як в сфері освіти першочерговим завданням є забезпечення якості освітнього процесу. Крім того, присутня суттєва відмінність в об'єктах інтересу та цільовій аудиторії рекомендаційних систем в освітній сфері та рекомендаційних систем комерційного призначення: чітко визначена аудиторія користувачів в освітній сфері на відміну від широкого кола користувачів рекомендаційних систем комерційного призначення; одноразовість та швидкоплинна актуальність багатьох рекомендацій в сфері освіти на відміну від можливості надавати однакову рекомендацію одним і тим же користувачам протягом тривалого часу в комерційних рекомендаційних

системах; можливість отримати попередні оцінки об’єктів інтересу за рахунок обов’язкових опитувань в сфері освіти на відміну від складного механізму отримання відгуків та оцінок від користувачів комерційних рекомендаційних систем. Такі суттєві відмінності потребують створення окремої методики для побудови рекомендаційної системи з метою підвищення якості вирішення різноманітних задач підтримки прийняття рішень в освітніх процесах.

В роботі вирішено актуальне науково-прикладне завдання підвищення якості надання рекомендацій в сфері освіти за рахунок використання рекомендаційної системи на основі гібридного підходу.

Проведено аналіз сучасних методів, моделей та підходів до побудови рекомендаційних систем, зокрема:

– підхід на основі контентної фільтрації (Content-Based Filtering), який використовує профілі користувачів та об'єктів на основі їх характеристик та рекомендує об'єкти, які подібні до тих, що користувач вже вибрав або виразив щодо них зацікавленість;

– підхід на основі колаборативної фільтрації (Collaborative Filtering), який враховує інформацію про взаємодію користувачів та їх історії вибору і заснований на припущенні, що користувачі, які мають спільні інтереси в минулому, будуть мати схожі інтереси у майбутньому;

– гібридні підходи, що комбінують різні методи для отримання більш точних рекомендацій а також для зменшення обсягів даних, що обробляються в процесі формування рекомендації.

Оскільки для побудови рекомендації враховуються результати оцінювань користувачами тих чи інших об’єктів інтересу, в роботі проведено детальний огляд методів, які використовуються в задачах експертного оцінювання. Розглянуто особливості визначення та специфіку використання номінальної шкали (шкали найменувань), порядкової (рангової) шкали, інтервальної (метричної) шкали та шкали відношення (абсолютної шкали). Для вирішення задачі визначення елементів шкал проведено аналіз підходу на основі методу ранжування, методу попарного порівняння та безпосередньої

(бальної) оцінки. Розглянуто питання формування узагальнених експертних оцінок для обчислення загального рейтингу об’єктів інтересу користувачів.

Завдяки тому, що значна частина характеристик як об’єктів інтересу так і параметрів запитів (пошуку) користувачів має лінгвістичну природу, було запропоновано використовувати модель представлення даних на основі нечіткої логіки. Перехід від числових множин елементів числових шкал до терм-множин дозволяє забезпечити спрощений перехід від однієї числової шкали до іншої при цьому не змінюючи лінгвістичне представлення моделі даних.

Застосування методів рекомендаційних систем в сфері освіти розглянуто в роботі в контексті найбільш типових задач здобувачів, що забезпечують підтримку освітнього процесу, а саме:

– рекомендація дисципліни з каталогу вибіркових дисциплін при формуванні індивідуальної освітньої траєкторії студента;

– рекомендація щодо можливості/доцільності використання того чи іншого джерела з бібліотеки закладу освіти при вивченні дисципліни чи для самоосвіти;

– рекомендація щодо вибору теми (або тематики дослідження) кваліфікаційної роботи чи наукової роботи, проєкту тощо;

– рекомендація при виборі потенційного керівника кваліфікаційної роботи, наукової роботи, проєкту тощо;

– рекомендація щодо використання зовнішніх курсів для організації самоосвіти (за умови, що відповідна інформація про ці курси присутня в базі рекомендаційної системи).

Розроблений метод формування рекомендацій по об’єктам інтересу з використанням гібридного підходу може бути використаний для зазначених та подібних до них рекомендацій з урахуванням наступних особливостей: системі відомий об’єкт інтересу користувача (для сфери вищої освіти – здобувача), об’єкти інтересу містять певний текстовий опис, в тому числі, з використанням неструктурованих текстів, які не обов’язково є відгуками, як в

класичних рекомендаційних системах; ступінь релевантності об’єкта інтересу до запиту користувача враховує в першу чергу семантичну близькість запиту (або характеристик користувача) до об’єктів інтересу, а вже на наступних кроках оцінки користувачів та інформацію про взаємодію здобувача з об’єктом інтересу та його поведінку в системі. Для оцінки релевантності інтересів користувача до наявних об’єктів рекомендації пропонується використовувати методи обробки природної мови. Застосування попередньої фільтрації об’єктів інтересу з використанням методів обробки текстової інформації дозволяє скоротити розмір вектору потенційно корисних об’єктів. Специфічні для освітнього процесу цільові опитування використовуються з тою ж метою, що дозволяє в подальшому суттєво зменшити розміри матриць колаборативної фільтрації на останньому етапі формування рекомендації. Також, рейтингування об’єктів інтересу на основі цільових опитувань дозволяє зменшити проблему «холодного старту», оскільки кількість об’єктів, про які немає рейтингової інформації в системі, є дуже низькою, і навіть для таких об’єктів можна легко отримати попередню оцінку на основі аналогічних оцінок схожих з ними об’єктів.

Апробація методики була проведена шляхом розробки рекомендаційної системи для вирішення задач формування рекомендації щодо вибору дисципліни блоку вільного вибору студента та підбору навчальної літератури. Аналіз отриманих результатів продемонстрував, що відгук моделі становить 92% - стільки користувачів погодились з рекомендацією, наданою системою. Оцінка рівня задоволеності користувачів отриманою рекомендацією щодо вибору дисципліни з плином часу становить 76%, при цьому 15% хоча і хотіли б поміняти свій вибір, але на ті дисципліни, що надала їх рекомендаційна система в вершині списку рекомендацій. Дослідження рівня релевантності рекомендації при вирішенні задачі підбору літературних джерел показало середню оцінку 8,5 за 10-бальною шкалою.

Ключові слова: рекомендаційна система; освітня сфера; автоматизація; інформаційна технологія; інформаційна система; підтримка прийняття рішень; математичне моделювання; база даних; нечітка логіка; нечітка модель; результати користувацької активності; машинне навчання; обробка текстів природною мовою; гібридний підхід.

ABSTRACT

Khudik B.O. Methodology for building a recommendation system for the field of education based on a hybrid approach. – Qualifying scientific work on the rights of the manuscript.

Dissertation for the Doctor of Philosophy degree in the specialty 123 Computer Engineering. - State University of Information and Communication Technologies of the Ministry of Education and Science of Ukraine. - Kyiv, 2023.

Recommender systems are a subclass of decision-making systems that use knowledge about a person's interests and preferences to evaluate and predict his reaction to a recommendation to buy a certain product or use a certain service. Modern approaches to the organization of educational processes in the formation of the individual educational trajectory of education seekers require the formation of certain recommendations regarding the selection of educational courses, educational literature, forecasting the enrollment of users for one or another course, and solving other auxiliary educational tasks, which makes these processes the sphere of interest of recommendation systems. Analysis of the area of use of recommender systems showed that most research and development is focused on ensuring commercial tasks and goals, which are related to increasing income, expanding the client base, etc., while in the field of education, the primary task is to ensure the quality of the educational process. In addition, there is a significant difference in the objects of interest and the target audience of recommender systems in the educational sphere and recommender systems for commercial purpose: a clearly defined audience of users in the educational sphere, in contrast to a wide range of users of recommender

systems for commercial purpose; the one-timeness and fleeting relevance of many recommendations in the field of education, in contrast to the ability to provide the same recommendation to the same users for a long time in commercial recommender systems; the ability to obtain preliminary evaluations of objects of interest due to mandatory surveys in the field of education, in contrast to the complex mechanism of receiving feedback and evaluations from users of commercial recommendation systems. Such significant differences require the creation of a separate methodology for building a recommendation system in order to improve the quality of solving various tasks of supporting decision-making in educational processes.

The work solves the actual scientific and applied task of improving the quality of providing recommendations in the field of education through the use of a recommendation system based on a hybrid approach.

An analysis of modern methods, models and approaches to the construction of recommendation systems was carried out, in particular:

– an approach based on content filtering, which uses profiles of users and objects based on their characteristics and recommends objects that are similar to those that the user has already chosen or expressed interest in;

– an approach based on collaborative filtering, which takes into account information about the interaction of users and their history of choices and is based on the assumption that users who have shared interests in the past will have similar interests in the future;

– hybrid approaches that combine different methods to obtain more accurate recommendations, as well as to reduce the amount of data processed in the process of forming a recommendation.

Since the results of user evaluations of certain objects of interest are taken into account for the construction of the recommendation, the work provides a detailed review of the methods used in expert evaluation tasks. The peculiarities of the definition and specifics of the use of the nominal scale (name scale), ordinal (rank) scale, interval (metric) scale and ratio scale (absolute scale) are considered. To solve the problem of determining the elements of the scales, an analysis of the approach

based on the ranking method, the pairwise comparison method, and direct (point) assessment was carried out. The issue of forming generalized expert assessments for calculating the general rating of objects of interest to users is considered.

Due to the fact that a significant part of the characteristics of both the objects of interest and the parameters of users' requests (search) has a linguistic nature, it was proposed to use a data presentation model based on fuzzy logic. The transition from numerical sets of elements of numerical scales to term-sets allows for a simplified transition from one numerical scale to another without changing the linguistic representation of the data model.

The application of the methods of recommendation systems in the field of education is considered in the work in the context of the most typical tasks of applicants that provide support for the educational process, namely:

– recommendation of a discipline from the catalog of selective disciplines when forming a student's individual educational trajectory;

– a recommendation regarding the possibility/expediency of using one or another source from the library of the educational institution when studying a discipline or for self-education;

– a recommendation regarding the choice of a topic (or research topic) of a qualification paper or research paper, project, etc.;

– recommendation when choosing a potential supervisor of a qualification work, scientific work, project, etc.;

– a recommendation on the use of external courses for the organization of self-education (provided that relevant information about these courses is available in the recommendation system database).

The developed method of forming recommendations on objects of interest using a hybrid approach can be used for the specified and similar recommendations, taking into account the following features: the system knows the object of the user's interest (for the field of higher education - the applicant), the objects of interest contain a certain text description, including, using unstructured texts, which are not necessarily reviews, as in classic recommender systems; the degree of relevance of

the object of interest to the user's request takes into account, first of all, the semantic proximity of the request (or user characteristics) to the objects of interest, and at the next steps of user evaluation and information about the interaction of the acquirer with the object of interest and his behavior in the system. To assess the relevance of the user's interests to the available objects, the recommendation suggests using natural language processing methods. Application of pre-filtering of objects of interest using text information processing methods allows to reduce the size of the vector of potentially useful objects. Targeted surveys specific to the educational process are used for the same purpose, which allows for a significant reduction in the size of collaborative filtering matrices at the final stage of recommendation formation. Also, the rating of objects of interest based on targeted surveys allows to reduce the problem of "cold start", since the number of objects about which there is no rating information in the system is very low, and even for such objects it is easy to get a preliminary assessment on based on similar assessments of similar objects.

Approbation of the methodology was carried out by developing a recommendation system to solve the problems of forming a recommendation regarding the choice of the discipline of the student's free choice block and the selection of educational literature. Analysis of the obtained results showed that the response of the model is 92% - so many users agreed with the recommendation provided by the system. The level of satisfaction of users with the received recommendation for choosing a discipline over time is 76%, while 15% would like to change their choice, but to those disciplines that the recommender system provided them at the top of the list of recommendations. The study of the level of relevance of the recommendation when solving the task of selecting literary sources showed an average rating of 8.5 on a 10-point scale.

Keywords*:* recommendation system; educational sphere; automation; information technology; information system; decision support; mathematical modeling; database; fuzzy logic; fuzzy model; user activity results; machine learning; natural language text processing; hybrid approach.

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА

Наукові праці, в яких опубліковані основні наукові результати дисертації:

1. Худік Б.О. Застосування рекомендаційної системи в освітній сфері. Зв'язок, № 6, 2023. С. 58-63. https://doi.org/10.31673/2412-9070.2023.065868

2. Худік Б.О. Модель представлення даних рекомендаційної системи в сфері освіти на основі нечіткої логіки. Том 1 № 21 (2023): Кібербезпека: освіта, наука, техніка. С. 260-272. https://doi.org/10.28925/2663-4023.2032.21.260272

3. Худік Б.О. Метод автоматизованого формування рекомендацій для здобувачів освіти на основі гібридного підходу. Телекомунікаційні та інформаційні технології. Київ, 2022, №4 (81). С. 21-27. https://doi.org/10.31673/2412-4338.2023.042127

Наукові праці, які засвідчують апробацію матеріалів дисертації:

1. Bohdan Khudik. Analysis of the approaches to the development of the multiagent recommender system based on the fuzzy logic// ITU Workshop for Europe and CIS Regions. ICT infrastructure as a basis for digital economy. 14-16 may 2019. International Telecommunication Union, State University of Telecommunications. P. 56.

2. Худік Б.О. Визначення рейтингу навчальної дисципліни на основі інтернет-опитувань // Міжнародна науково-практична конференція «Сучасні аспекти діджиталізації та інформатизації в програмній та комп'ютерній інженерії». Збірник тез. К.: ДУТ, 2023. С. 167-169.

3. Різанова С.Д., Худік Б.О. Інтеграція штучного інтелекту для оптимізації процесів розробки веб-дизайну, інтерфейсів та супутніх аспектів в цифровому просторі// Всеукраїнська науково-технічна конференція «Технологічні горизонти: дослідження та застосування інформаційних технологій для технологічного прогресу України і Світу». Збірник тез. К.: ДУІКТ, 2023. С. 160-161.

4. Худік Б.О. Використання нечітких моделей даних в рекомендаційних системах// Науково-практична конференція «Проблеми комп’ютерної інженерії». Збірник тез. К.: ДУІКТ, 2023. С. 154-155.

5. Худік Б.О. Вилучення інформації з освітніх джерел на основі Natural Language Processing// «ТАК»: телекомунікації, автоматика, комп’ютерно-інтегровані технології: зб. доповідей Всеукр. наук.-практ. конф. молодих вчених, 5-6 грудня 2023 р./ ДВНЗ «ДонНТУ»; відп. ред. Г.В. Ступак. Луцьк: ДВНЗ «ДонНТУ», 2023. С. 27-29.

ЗМІСТ

ВСТУП 14
РОЗДІЛ 1 АНАЛІЗ МЕТОДІВ, МОДЕЛЕЙ ТА ПІДХОДІВ ДО ПОБУДОВИ РЕКОМЕНДАЦІЙНИХ СИСТЕМ..... 21
1.1 Загальна постановка задач рекомендаційної системи..... 21
1.2 Аналіз підходів до побудови рекомендаційних систем 26
1.2.1 Використання методів колаборативної фільтрації для побудови рекомендацій..... 28
1.2.3 Контентний підхід до побудови рекомендацій 30
1.2.4 Гібридні рекомендаційні системи 32
1.3 Аналіз інфокомунікаційних засобів для забезпечення підтримки процесів в рекомендаційних системах..... 36
1.4 Постановка наукового завдання та визначення мети дослідження 41
Висновки до розділу 1 42
РОЗДІЛ 2. АНАЛІЗ МОДЕЛЕЙ ТА МЕТОДІВ ОБРОБКИ ДАНИХ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ 43
2.1 Моделі та методи обробки експертних даних для задачі колаборативної фільтрації..... 43
2.2 Моделі та методи обробки даних рекомендаційної системи на основі нечіткої логіки 54
2.3 Підходи до обробки текстових даних природною мовою 65
2.3.1 Методи попередньої обробки даних та підходи до векторизації даних. 66
2.3.2 Методи Natural Language Processing для визначення ключових елементів тексту та семантичних зв’язків..... 74
2.3.4 Моделі та методи обробки текстових даних на основі нейронних мереж75
Висновки до розділу 2 81
РОЗДІЛ 3. РОЗРОБКА МЕТОДИКИ ПОБУДОВИ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ ДЛЯ СФЕРИ ОСВІТИ НА ОСНОВІ ГІБРИДНОГО ПІДХОДУ . 83
3.1 Визначення характеристик рекомендаційної системи для сфери освіти .. 83

3.1.1 Визначення задач користувачів рекомендаційної системи для сфери освіти 83
3.1.2 Визначення характеристик об'єктів інтересу 85
3.1.3 Визначення характеристик користувачів рекомендаційної системи..... 90
3.2 Моделювання структури рекомендаційної системи..... 92
3.3 Розробка моделі даних рекомендаційної системи для сфери освіти на основі нечіткої логіки 95
3.4 Розробка методу формування рекомендацій по об’єктам інтересу з використанням гібридного підходу..... 105
3.5 Алгоритм застосування методики побудови рекомендаційної системи для сфери освіти на основі гібридного підходу..... 108
Висновки по розділу 3 112
РОЗДІЛ 4. ПРАКТИЧНА РЕАЛІЗАЦІЯ ТА ОЦІНКА ЕФЕКТИВНОСТІ МЕТОДИКИ ПОБУДОВИ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ ДЛЯ СФЕРИ ОСВІТИ НА ОСНОВІ ГІБРИДНОГО ПІДХОДУ..... 114
4.1 Визначення параметрів моделей та методів рекомендаційної системи для формування рекомендації навчальної дисципліни та літератури 114
4.2 Розробка структур баз даних рекомендаційної системи для формування рекомендації навчальної дисципліни та літератури 126
4.3 Моделювання архітектури та реалізація рекомендаційної системи для формування рекомендації навчальної дисципліни та літератури 131
4.4 Оцінка ефективності розробленої методики 134
Висновки по розділу 4 139
Висновки 140
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ..... 143
ДОДАТОК А..... 161
ДОДАТОК Б. 163
ДОДАТОК В. 164

ВСТУП

Актуальність теми. Однією з важливих задач в сфері вищої освіти є забезпечення необхідних показників якості надання освітніх послуг, що досягається в тому числі за рахунок всебічної підтримки освітніх процесів засобами комп'ютерних систем та інформаційних технологій. В умовах широкого запровадження технологій дистанційного навчання стає актуальним питання організації навчального процесу таким чином, щоб здобувач мав високий рівень інформаційної, організаційної, методичної та інших видів підтримки не зважаючи на зменшення можливості чи повну неможливість отримати консультаційну допомогу від адміністрації чи викладачів закладу освіти. Забезпечення такого роду підтримки може бути реалізовано різними програмними та програмно-апаратними засобами, в тому числі, за допомогою рекомендаційних систем. Рекомендаційні системи ефективно впроваджуються там, де потрібно забезпечити персоналізовані та адаптовані під конкретного користувача рекомендації з великого обсягу інформації [1]. Зважаючи на те, що в загальному випадку рекомендаційна система може бути використана для визначення цікавих для користувача об'єктів будь-якого характеру, в тому числі, не пов'язаних безпосередньо товарно-грошовими операціями (операція покупки товару чи використання послуги) [2], їх використання в освітній сфері є цілком доцільним.

Значна частина досліджень та практичних розробок в галузі рекомендаційних систем зосереджується на комерційних задачах для стрімінгових сервісів, новинних ресурсів, соціальних мереж, веб-ресурсів електронної комерції, освітніх платформ в секторі інформальної освіти, рекомендацій літератури, продукції харчової індустрії, медичної сфери та ігрової індустрії [3, 4, 5, 6, 7, 8]. При цьому, підходи до побудови рекомендаційних системи, методи обробки інформації про користувачів та об'єкти їх інтересу, методи та засоби формування рекомендації, способи та

метрики оцінки якості наданих рекомендацій обов’язково адаптуються відповідно до цілей та задач впровадження такої системи [9, 10].

Загалом, розрізняють декілька груп методів, що використовуються в рекомендаційних системах. В першу чергу, це колаборативна, або сумісна, фільтрація, методи, моделі та алгоритми якої ґрунтується на даних про попередні сеанси роботи користувачів з системою [11]. Не зважаючи на широке використання колаборативної фільтрації для побудови різноманітних рекомендаційних систем, вони мають багато серйозних проблем, що в сучасних умовах ставить під сумнів доцільність їх використання в процесі формування якісної рекомендації. Зокрема, ряд досліджень констатують наступні проблеми: холодний старт, при якому важко отримати адекватну рекомендацію для нового користувача чи нового продукту, проблема розрідженості матриці взаємодії між товарами та користувачами, проблема поганих оцінок, що зменшує точність моделювання вподобань користувачів, проблеми масштабування при великій кількості об’єктів інтересу та користувачів, проблема визначення схожості користувачів особливо для нових або малочисельних користувачів [1, 12, 13]. Альтернативою для колаборативної фільтрації виступають методи, засновані на контентній фільтрації, яка спирається на дані та метадані об’єктів, а в деяких випадках, і профілів користувачів. Тим не менш, контентні методи мають ряд суттєвих недоліків, які включають наступні: проблема обчислювальної складності, що призводить до суттєвого зниження продуктивності системи, проблема недостатньої інформації , проблема персоналізації (оскільки контентні методи в першу чергу грунтуються на аналізі характеристик об’єкту, а не враховують особисті вподобання користувача), проблема недостатності опису контенту та інші, що не також ускладнюють їх використання для отримання якісної рекомендації [1, 14]. Сучасним рішенням для побудови рекомендаційних систем є гібридні моделі, які за рахунок комбінації різних методів та моделей представлення та обробки даних дозволяють забезпечити суттєве підвищення якості наданої рекомендації [15, 16].

Таким чином, з урахуванням нагальних потреб підвищення якості процесів в сфері освіти, які стосуються автоматизованої підтримки прийняття рішень при вирішенні різноманітних завдань, виходячи з оцінки існуючих моделей, методів та підходів до побудови рекомендацій, можна зробити висновок про те, що на сьогоднішній день актуальним науковим завданням є розробка методики побудови рекомендаційної системи для сфери освіти на основі гібридного підходу.

Мета та завдання дослідження.

Метою дослідження є підвищення якості надання рекомендацій в сфері освіти за рахунок використання рекомендаційної системи на основі гібридного підходу з використанням нечіткої логіки та методів обробки природньої мови.

Для досягнення поставленої мети в дисертаційній роботі поставлені та вирішені наступні завдання:

1. Провести аналіз моделей та методів обробки даних рекомендаційних систем.
2. Визначити характеристики рекомендаційної системи для сфери освіти.
3. Розробити модель представлення даних рекомендаційної системи для освітньої сфери на основі нечіткої логіки.
4. Провести моделювання структури рекомендаційної системи для сфери освіти.
5. Розробити метод формування рекомендації по об’єктам інтересу освітньої сфери на основі гібридного підходу.
6. Розробити алгоритм застосування методики побудови рекомендаційної системи для сфери освіти на основі гібридного підходу.
7. Визначити параметри моделей та методів рекомендаційної системи, розробити структури баз даних, провести моделювання архітектури системи та розробити програмні засоби для задачі формування рекомендації в сфері освіти.

8. Провести дослідження результатів застосування розробленої методики побудови рекомендаційної системи в сфері освіти на основі гібридного підходу.

Об'єктом дослідження є процес формування рекомендації в освітній сфері.

Предметом дослідження є моделі та методи формування рекомендації в освітній сфері з використання гібридного підходу, що ґрунтується на нечіткій логіці та методах обробки природної мови.

Методи дослідження: методи обробки експертних даних; методи обробки даних на основі нечіткої логіки; методи штучного інтелекту; методи колаборативної фільтрації; методи контентної фільтрації; гібридні методи рекомендаційних систем; методи обробки текстових даних природною мовою; методи кластеризації; експериментальні дослідження; методи оцінки ефективності рекомендаційних систем.

Наукова новизна одержаних результатів.

1. Розроблено модель представлення даних рекомендаційної системи для освітньої сфери на основі нечіткої логіки, яка, на відміну від існуючих, дозволяє використовувати лінгвістичну форму в процесі експертних оцінок параметрів.

2. Розроблено структурну модель рекомендаційної системи для освітньої сфери, яка, на відміну від існуючих, дозволяє формувати комплексну рекомендацію з урахуванням наявних інформаційних ресурсів.

3. Розроблено метод формування рекомендацій на основі гібридного підходу, який, на відміну від існуючих, ґрунтується на нечітких моделях представлення та обробки даних про об'єкти інтересу та користувачів і методах обробки природної мови, що дозволяє підвищити точність рекомендацій.

4. Розроблено методику побудови рекомендаційної системи для сфери освіти на основі гібридного підходу, яка включає структурну модель рекомендаційної системи та метод формування рекомендацій, що ґрунтується

на нечітких моделях представлення та обробки даних про об’єкти інтересу та користувачів і методах обробки природньої мови.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконана відповідно до планів наукових досліджень в Державному університеті інформаційно-комунікаційних технологій (м. Київ) в рамках науково-дослідної роботи «Розробка архітектури програмного комплексу машинного навчання для IoT-пристроїв»» (ДР № 0121U111079), в якій автором розроблено модель представлення даних рекомендаційної системи для освітньої сфери на основі нечіткої логіки, структурну модель рекомендаційної системи для освітньої сфери, метод формування рекомендацій на основі гібридного підходу, який ґрунтується на нечітких моделях представлення та обробки даних про об’єкти інтересу та користувачів і методах обробки природної мови, та узагальнено зазначені розробки у вигляді методики побудови рекомендаційної системи для сфери освіти на основі гібридного підходу.

Практичне значення одержаних результатів.

Результати дослідження можуть бути використані при побудові рекомендаційних систем в освітній сфері для вирішення задачі якісної підтримки освітнього процесу шляхом перенесення частини функцій персоналу закладу освіти в рекомендаційну систему при наданні консультаційних та рекомендаційних послуг.

Застосування запропонованої в дисертаційній роботі методики побудови рекомендаційної системи в сфері освіти на основі гібридного підходу при вирішенні задачі закріплення здобувачів за дисциплінами вільного вибору продемонструвало збіг рекомендації з вибором здобувачів освіти у 92% випадків. При вирішенні задачі підбору літературних джерел оцінка релевантності запропонованих рекомендаційною системою варіантів до реальних потреб користувачів склала в середньому 8,5 за 10-бальною шкалою.

Наукові положення, теоретичні та експериментальні результати, отримані в дисертаційній роботі, використовуються в навчальному процесі

Державного університету інформаційно-комунікаційних технологій при курсовому проектуванні, а також при викладанні навчальних дисциплін: «Технології проектування та адміністрування баз даних та сховищ даних», «Технології проектування комп'ютерних систем», «Моделювання комп'ютерних систем», «Інтелектуальна обробка даних».

Особистий внесок здобувача. Основні наукові та практичні результати дисертаційної роботи отримані автором особисто.

Апробація результатів дисертації. Результати досліджень, викладених у дисертаційній роботі, доповідались та обговорювались на наступних конференціях: ITU Workshop for Europe and CIS Regions. ICT infrastructure as a basis for digital economy (14-16 травня 2019 р., International Telecommunication Union, Державний університет телекомунікацій, м. Київ), Міжнародна науково-практична конференція «Сучасні аспекти діджиталізації та інформатизації в програмній та комп’ютерній інженерії» (01-03 червня 2023 р., Державний університет телекомунікацій, м. Київ), Всеукраїнська науково-технічна конференція «Технологічні горизонти: дослідження та застосування інформаційних технологій для технологічного прогресу України і Світу» (28 листопада, Державний університет інформаційно-комунікаційних технологій, м. Київ), Науково-практична конференція «Проблеми комп’ютерної інженерії» (1 грудня, Державний університет інформаційно-комунікаційних технологій, м. Київ), Всеукраїнська науково-практична конференція молодих вчених «ТАК»: телекомунікації, автоматика, комп’ютерно-інтегровані технології» (5-6 грудня 2023 р., ДВНЗ «Донецький національний технічний університет», м. Луцьк).

Публікації. Основні положення та результати дисертаційної роботи опубліковано в 8 друкованих працях, з них: 3 статті, опубліковані у наукових фахових виданнях України, затверджених МОН України [17-19], і 5 публікацій у збірниках матеріалів науково-практичних та науково-технічних конференцій [20-24].

Структура та обсяг дисертації. Дисертація складається із вступу, 4 розділів, висновків, списку використаних джерел із 166 найменувань та 3 додатки.

Перший розділ присвячено аналізу методів, моделей та підходів до побудови рекомендаційних систем.

В другому розділі проведено аналіз моделей та методів обробки даних рекомендаційної системи.

Третій розділ містить розробку методики побудови рекомендаційної системи для сфери освіти на основі гібридного підходу.

Четвертий розділ демонструє практичну реалізацію рекомендаційної системи та оцінку ефективності методики побудови рекомендаційної системи для сфери освіти на основі гібридного підходу.

Загальний обсяг дисертації становить 168 сторінок, серед яких 24 рисунки та 12 таблиць.

РОЗДІЛ 1
АНАЛІЗ МЕТОДІВ, МОДЕЛЕЙ ТА ПІДХОДІВ ДО ПОБУДОВИ РЕКОМЕНДАЦІЙНИХ СИСТЕМ

1.1 Загальна постановка задач рекомендаційної системи

Сьогодні рекомендаційні системи є корисним інструментом підтримки прийняття рішень, а також невід'ємною частиною повсякденного життя будь-якого користувача та веб-сервісів. Основними складовими рекомендаційні системи є користувач і елемент. Численні алгоритми, описані для рекомендаційні системи, ґрунтуються на зворотному зв'язку, наданому користувачами щодо елемента у вигляді огляду, тегів і рейтингу, або відстежують поведінку користувача з точки зору його вподобань і антипатій щодо елементів на основі цих інформаційних алгоритмів, які дають прогнози.

Рекомендаційні системи знайшли широке розповсюдження в сфері електронної комерції, соціальних мережах, що обумовлено їх доволі високою ефективністю при формуванні пропозицій для потенційних покупців/клієнтів компаній чи побудові рекламних кампаній або вибудові цифрових зв'язків між користувачами [25].

Рекомендаційні системи використовується для обробки великої кількості даних, щоб рекомендувати новий товар користувачам, що досягається за допомогою багатьох розроблених алгоритмів. Для багатьох онлайн-додатків є складним завданням створити ефективний алгоритм для рекомендаційних систем, який дотримується гарного компромісу між точністю та різноманітністю. Різноманітність у рекомендаційних системах використовується для уникнення проблеми переобладнання, а також відмінні навички, які забезпечують рекомендації на основі підвищення якості взаємодії з користувачем. Вплив різноманітності на систему сприяє покращенню взаємодії з користувачами з новими елементами. [26]

Як програмний інструмент, рекомендаційні системи призначені для аналізу минулого досвіду користувача та надають список пропозицій із

великої кількості інформації. Багато алгоритмів рекомендацій розроблено, щоб дізнатися про минулу поведінку користувача після того, як ця рекомендація згенерована відповідно до історії їхніх переваг [27]. Рекомендаційна система — це також техніка, яка використовується для надання пропозицій користувачеві щодо вибору товару. Ці пропозиції ґрунтуються на різних процесах прийняття рішень, тобто вибір предметів для покупки, перегляд фільму з набору фільмів, подібно до інших онлайн-додатків [28]. Потрібно відмітити, що різні типи алгоритмів рекомендацій працюють відповідно до їх відповідної області та знань, які використовуються для користувачів, а для генерації рекомендацій використовуються різні типи алгоритмів прогнозування. Існує багато алгоритмів, визначених як спільна фільтрація, рекомендаційна система на основі вмісту та гібридна рекомендаційна система. Алгоритм спільної фільтрації заснований на фільтрації інформації або знаходженні взаємопов’язаних шаблонів за допомогою різних методів, що включають співпрацю між різними користувачами та елементами [29]. Іншим методом, окрім спільної фільтрації, є алгоритми на основі вмісту, коли система намагається рекомендувати користувачам елементи, подібні до попередніх уподобань іншого користувача [30]. Подібність між елементами та користувачами обчислюється за допомогою різних показників подібності. Подібність між різними користувачами та елементом обчислюється за допомогою інформації про вміст, наданої користувачем. Як спільна фільтрація, так і підхід на основі вмісту мають свої плюси та мінуси. Підхід спільного фільтрування страждає від проблеми, коли для нового елемента та генерування нових рекомендацій для користувачів неможливо, подібно до генерації рекомендацій на основі вмісту, специфікація конкретного опису вмісту викликає плутанину. З іншого боку, підхід спільної фільтрації також має проблему від розрідженості, що означає, що наявна кількість елементів перевищує ту кількість, яку людина може дослідити. У техніках фільтрації на основі контенту іноді виникають труднощі з розрізненням особистої інформації користувача [31]. У минулому

оцінка в системі рекомендацій залежала від точності, що означає, наскільки релевантні елементи рекомендуються користувачеві [32]. Але в наш час для надійної рекомендаційної системи використовується занадто багато інших оціночних заходів: різноманітність, випадковість і новизна.

Дослідженням питання функціонування та застосуванням рекомендаційних систем проводились у роботах Patel Dhruval, Patel Foram, Chauhan Uttam, Мелешко Є. В., Beheshti A., Yakhchi, S., Mousaeirad S., Ghafari S. M., Goluguri S. R., Edrisi M. A., Raghuwanshi Sandeep, Pateriya R., Yuan Zhenning, Lee Jong та Zhang Sai.

Робота [9] присвячена огляду рекомендаційних систем, їх типів, особливостей застосування та типових проблем побудови і використання. Розглядаються методи, які використовуються в рекомендаційних системах, включаючи фільтрацію на основі вмісту, колаборативну (соціальну) фільтрацію, послідовну, сеансову фільтрацію та інші. Наведена загальна таксономія методів, що використовуються в рекомендаційних системах, а також метрики оцінки їх ефективності. Автори проводять аналіз наборів даних та особливостей доменів декількох найбільш популярних предметних галузей використання рекомендаційних систем, серед яких в якості ТОП-5 галузей виділено фільми, книги, їжа, музика та деякі продукти чи послуги. Доводиться корисність рекомендаційних систем в боротьбі з інформаційним шумом, що супроводжує користувачів різноманітних Інтернет-систем. Розглядаються різноманітні проблеми рекомендаційних систем та пропонуються деякі рішення для їх усунення з метою підвищення точності надання рекомендації.

В роботі [1] досліджуються проблеми сучасних рекомендаційних систем, в тому числі, пов’язані з холодним стартом – ситуацією, в якій не накопичено достатньої кількості даних для коректної роботи системи. Для усунення цієї проблеми рекомендується використовувати гібридний підхід до побудови рекомендації, який враховує контекстну інформацію.

Конкретні методи та підходи до формування рекомендацій наведено в роботах [33-35].

Інтелектуальна система рекомендацій, яка розуміє уподобання користувача, виявляє зміни в уподобаннях користувача з часом, передбачає невідомі уподобання користувача та досліджує адаптивні механізми для забезпечення інтелектуальних дій у складному та змінному середовищі описується в роботі [33]. В результаті аналізу впровадження рекомендаційної системи в банківській сфері автори доходять до висновку, що існуючі рекомендаційні системи мало або зовсім не використовують знання експертів галузі в процесі адаптації до нової ситуації, часто не в змозі адекватно передбачити переваги того чи іншого продукту, який обрав користувач/клієнт, не підтримують збирання даних та аналітику діяльності користувачів і не використовують ці дані для надання оперативних рекомендацій. Запропонована в статті система когнітивних рекомендацій характеризується тим, що вона керується знаннями, дозволяє імітувати роботу експертів на основі методів краудсорсингу, використовує технології штучного інтелекту та машинного навчання для розуміння великих даних та дозволяє зрозуміти особистість кінцевих користувачів і проаналізувати їхню поведінку та ставлення з протягом деякого часу.

В роботі [34] розглянуто один із найбільш популярних методів реалізації рекомендаційних систем – метод колаборативної фільтрації та його реалізації: фільтрація на основі класів, на основі пам’яті та на основі моделі, наводяться приклади реалізації різновидів методів та їх порівняльна характеристика. Суттєвою проблемою рекомендаційних систем автори вважають різноманітність предметних галузей, що ускладнює або унеможливлює точний прогноз з використанням однієї й тієї ж техніки побудови рекомендації для різних програм.

В статті [35] пропонується гібридна система рекомендацій фільмів і метод оптимізації, заснований на зваженій класифікації та алгоритмі колаборативної фільтрації користувачів. Розріджена лінійна модель використовується як базова модель рекомендацій, локальна модель рекомендацій навчається на основі кластеризації користувачів, а

персоніфікована рекомендація фільмів із топ-N реалізується шляхом об’єднання з моделлю зваженої класифікації. Згідно з уподобаннями категорії предметів, оціночна матриця перетворюється на низькорозмірну, щільну матрицю переваг категорії предметів, отримується кілька центрів кластерів, обчислюється відстань між цільовим користувачем і кожним центром кластеру, а цільовий користувач класифікується на найближчий кластер. Після цього алгоритм колаборативної фільтрації використовується для прогнозування балів товарів без оцінки цільового користувача для формування списку рекомендацій. Елементи кластеризуються за допомогою параметра категорії елемента, а багатовимірна рейтингова матриця перетворюється на низьковимірну матрицю переваг категорії елемента, що додатково зменшує розрідженість даних.

Питанням використання рекомендаційних систем в галузі освіти присвячено роботу [7]. Розглядається задача підтримки користувача у виборі навчального курсу в системі електронного навчання. Для формування рекомендації пропонується використовувати три основні методи: рекомендації на основі вмісту, на основі колаборативної фільтрації та на основі знань.

В загальному потрібно виділити, що основна постановка задачі рекомендаційної системи включає такі етапи:

1. Збирання і акумуляція різноманітних даних про користувачів (попередні вибори, перегляди, оцінки, демографічні дані тощо) та об'єкти (атрибути, характеристики, описи тощо).

2. Підготовка та очищення даних для подальшого аналізу. Це включає фільтрацію непотрібних даних, нормалізацію, кодування категорійних змінних тощо.

3. Вибір підходящого алгоритму (колаборативний, контентний, гібридний), який найкраще підходить для конкретного випадку за умовами та характеристиками даних.

4. Створення моделі рекомендаційної системи на основі вибраного методу, використовуючи алгоритми машинного навчання або інші методи аналізу даних.

5. Оцінка ефективності рекомендаційної системи за допомогою метрик якості, таких як точність, покриття, різноманітність рекомендацій тощо.

6. Впровадження системи в реальному середовищі та постійна підтримка, включаючи моніторинг та оновлення моделі на основі нових даних і змін у поведінці користувачів.

Дана постановка задач допомагає створити ефективну та корисну рекомендаційну систему, яка відповідає потребам користувачів та допомагає в різних сферах, від електронної комерції до освітнього контенту.

1.2 Аналіз підходів до побудови рекомендаційних систем

Більшість досліджень рекомендаційних систем зосереджуються на побудові рекомендацій та прогнозуванні вибору користувача відносно об’єктів комерційного спрямування, як от фільми, товари, музика тощо [9, 10, 11, 33]. Частина досліджень в галузі розробки рекомендаційних систем лежить в площині, дотичній до освітньої сфери. Наприклад, рекомендаційна система в роботі [35] дозволяє сформувати список рекомендованих для користувача книг на основі вивчення його вподобань та враховує оцінку книг та улюблені категорії. Основний акцент зроблено на швидкості формування рекомендації та відсутності потреби у зборі довготривалої статистики про роботу користувача на сайті. В роботі [7] розглядається питання побудови рекомендаційних систем в сфері E-Learning на основі домену, який цікавить користувачів. В якості одного з критеріїв якості рекомендації розглядається мінімальне відхилення в темах, що пропонує рекомендаційна система, на відміну від більш типового підходу, що ґрунтується на аналізі рейтингів, популярності курсів тощо.

Для безпосередньої побудови рекомендацій авторами використовуються різні підходи. Контентні рекомендаційні системи максимально використовують наявну інформацію про об’єкт, який хоч раз зацікавив користувача, чи у випадках, коли користувач намагається знайти схожий об’єкт [6]. Системи колаборативної фільтрації (соціальні) будують рекомендації на основі лише аналізу історії оцінок, при цьому намагаючись виявити певні приховані залежності [11]. Гібридні рекомендаційні системи поєднують властивості контентних та колаборативної фільтрації і намагаються врахувати всі доступні види інформації про об’єкт та користувача і історію його попередньої взаємодії з цим чи іншими об’єктами [35]. Моделі рекомендації при цьому можуть використовувати додаткові засоби обробки інформації про об’єкт чи користувача. Так, в роботі [36] рекомендація формується на основі колаборативної фільтрації разом із методами навчання на основі графів, такими як Gensim Word2vec і GraphSage. Для колаборативної фільтрації будується кореляційна матриця та зведена таблиця, щоб отримати зв’язок між користувачами і продуктами та виявлення зв’язків між різними користувачами, а в моделях навчання графів використовуються ембедінги слів, представлені графами, що створюються на основі зібраних даних. Оцінка кожної моделі виконується на основі методу персоналізації.

У відповідності до проведеного аналізу наукових джерел виділяють декілька підходів до побудови рекомендаційних систем.

Фільтрація за змістом (Content-Based Filtering). Цей підхід використовує характеристики або властивості елементів (такі як ключові слова, опис, категорії тощо), щоб рекомендувати подібні елементи.

Фільтрація на основі спільноти (Collaborative Filtering). Цей підхід використовує інформацію про поведінку користувачів і їхні взаємодії з елементами. Він може використовувати спільність між користувачами або елементами для рекомендацій. Наприклад, якщо користувач А має схожі

вподобання з користувачем B, система може рекомендувати елементи, які сподобалися користувачу B, користувачу A.

Гібридні системи (Hybrid Recommender Systems). Це поєднання різних підходів для забезпечення кращої якості рекомендацій. Гібридні системи можуть об'єднувати фільтрацію за змістом і фільтрацію на основі спільноти для надання більш точних та різноманітних рекомендацій.

Розгортання на основі знань (Knowledge-Based Systems). Цей підхід використовує додаткові знання або правила, які моделюють уявлення про те, що користувачі вважають важливим або що є правильним. Наприклад, якщо користувач шукає книги на певну тему, система може використовувати експертні знання або правила для рекомендацій.

Факторизація матриці (Matrix Factorization). Цей підхід базується на розкладанні матриць оцінок користувачів і елементів, що дозволяє зменшити розмірність та знайти складові, які краще описують взаємодії користувачів із елементами.

Заснований на глибокому навчанні (Deep Learning-Based): Використовуючи нейронні мережі з великою кількістю шарів для аналізу складних взаємодій між користувачами та елементами, цей підхід може навчати модель рекомендацій за допомогою глибокого навчання.

Всі ці підходи можуть використовуватися окремо або комбінуватися для створення більш точних та ефективних рекомендаційних систем.

1.2.1 Використання методів колаборативної фільтрації для побудови рекомендацій

Колаборативна фільтрація є одним із добре відомих і найпоширеніших методів у рекомендаційній системі, її основна ідея полягає в тому, щоб передбачити, які елементи зацікавлять користувача на основі його переваг. Рекомендаційні системи, що використовують колаборативну фільтрацію, здатні забезпечити точний прогноз, коли надається достатньо даних, оскільки

ця техніка базується на перевагах користувача. Колаборативна фільтрація на основі користувачів була дуже успішною в минулому для прогнозування поведінки клієнта як найважливішої частини рекомендаційної системи. Однак їх широке використання виявило деякі реальні проблеми, такі як розрідженість даних і масштабованість даних із поступовим збільшенням кількості користувачів і елементів.

Автоматизовані системи колаборативної фільтрації (ACF) передбачають приналежність користувача до елементів або інформації. На відміну від традиційної системи фільтрації інформації на основі вмісту, розробленої з використанням технології пошуку інформації або штучного інтелекту, рішення щодо фільтрації в ACF базуються на людському, а не машинному аналізі вмісту. Кожен користувач системи ACF оцінює елементи, з якими він стикався, щоб створити профіль інтересів. Потім система ACF об’єднує цього користувача з людьми зі схожими інтересами чи смаками [37]. Потім оцінки цих схожих людей використовуються для створення рекомендацій для користувача.

Система ACF має багато суттєвих переваг перед традиційною фільтрацією на основі вмісту, насамперед тому, що вона не залежить від машинного аналізу вмісту, схильного до помилок. До переваг можна віднести можливість фільтрації будь-якого типу вмісту, напр. текст, художній твір, музика, пайові фонди; можливість фільтрації на основі складних і складних для представлення понять, таких як смак і якість; і здатність давати випадкові рекомендації.

Важливо зазначити, що технології ACF не обов’язково конкурують із фільтрацією на основі вмісту. У більшості випадків їх можна інтегрувати, щоб забезпечити потужне гібридне рішення фільтрації.

Хоча автоматизовані системи спільної фільтрації виявилися достатньо точними для розважальних доменів, вони ще не досягли успіху в контентних доменах, де більший ризик пов’язаний із прийняттям рекомендацій щодо фільтрації.

В роботі [37] наведено причини, чому системам ACF не можна довіряти для доменів вмісту високого ризику.

По-перше, системи ACF — це стохастичні процеси, які обчислюють прогнози на основі моделей, які є евристичними наближеннями людських процесів. По-друге, і, ймовірно, найважливіше, системи ACF базують свої обчислення на надзвичайно розріджених і неповних даних. Ці дві умови призводять до рекомендацій, які часто є правильними, але іноді дуже неправильними. Сучасні системи ACF — це чорні скриньки, комп’ютеризовані оракули, які дають поради, але не можуть ставитися під сумнів. Користувачеві не надається жодних індикаторів, з якими можна ознайомитися, щоб визначити, коли довіряти рекомендації, а коли сумніватися. Ці проблеми запобігли прийняттю систем ACF у всіх доменах вмісту, крім низьких ризиків.

1.2.3 Контентний підхід до побудови рекомендацій

Контентний підхід до побудови рекомендацій базується на аналізі властивостей самого контенту об'єктів та врахуванні вподобань користувачів на основі цих властивостей. Він використовує характеристики об'єктів і зв'язки між ними для створення рекомендацій. В наукових працях багато авторів зосереджуються на контентному підході до побудови рекомендаційних систем, серед яких D. Billsus, M. J. Pazzani [38], C. Li, M. Zhang, Y. Zhang [39], S. Brin [40].

Так, в роботі [38] розглядаються системи рекомендацій на основі контенту, тобто системи, які рекомендують товар користувачеві на основі опису елемента та профілю інтересів користувача. Системи рекомендацій на основі вмісту можна використовувати в різних сферах, починаючи від рекомендацій веб-сторінок, новинних статей, ресторанів, телевізійних програм і товарів для продажу. Незважаючи на те, що деталі різних систем відрізняються, системи рекомендацій на основі вмісту мають спільні засоби для опису елементів, які можуть бути рекомендовані, засоби для створення

профілю користувача, який описує типи елементів, які подобаються користувачеві, і засоби порівняння елементів із профілем користувача, щоб визначити, що рекомендувати. Профіль часто створюється й оновлюється автоматично у відповідь на відгук про бажаність елементів, які були представлені користувачеві.

В роботі [39] досліджуються методи витягування інформації з текстів для покращення контентних систем рекомендацій, а в [40] розглядають застосування методів машинного навчання для ранжування контенту в системах рекомендацій.

Розгляд даних джерел дозволяє виділити основні переваги та недоліки контентного підходу до побудови рекомендацій. До переваг можна віднести:

1. Незалежність від користувачів. Не потрібно знати відгуки інших користувачів.

2. Легка інтерпретація. Рекомендації можна пояснити на основі характеристик об'єктів і вподобань користувачів.

3. Робота з новими об'єктами. Здатність працювати з новими об'єктами без історії взаємодії.

Щодо недоліків потрібно виділити:

1. Обмежена рекомендаційна точність. Може бути обмежена у точності порівняно з коллаборативними підходами.

2. Проблема фільтрації бульбашок. Може обмежувати рекомендації до вже відомого контенту.

3. Залежність від якості опису контенту. Якість рекомендацій залежить від якості та повноти опису об'єктів.

Цей підхід до побудови рекомендацій широко використовується, особливо там, де існує багатий контент з визначеними характеристиками, такими як книги, фільми, музика та товари.

1.2.4 Гібридні рекомендаційні системи

Технічно всі рекомендаційні системи використовують одну або більше стратегій рекомендацій, таких як фільтрація на основі вмісту (CBF), фільтрація на основі спільної роботи (CF), фільтрація за демографічними ознаками (DF) або фільтрація на основі знань (KBF).

Побудова рекомендаційної системи на основі колаборативної фільтрації використовує рейтинги або інші форми зворотного зв’язку, створеного користувачами, щоб виявити спільні смаки між групами користувачів, а потім генерує рекомендації на основі схожості між користувачами [41].

В свою чергу побудова рекомендаційної системи з використанням підходу фільтрування на основі вмісту базується на припущенні, що люди, яким подобалися предмети з певними атрибутами в минулому, подобатимуться такого ж типу предметів у майбутньому. Він використовує функції предмета для порівняння товару з профілями користувачів і надання рекомендацій. Якість рекомендацій обмежена вибраними функціями рекомендованих елементів. Даний підхід страждає від проблеми холодного запуску.

Іншим підхід побудови рекомендаційної системи, який базується на демографічному фільтруванні, коли використовується демографічні дані, такі як вік, стать, освіта тощо, для визначення категорій користувачів. Він не страждає від проблеми нового користувача, оскільки не використовує рейтинги для надання рекомендацій. Однак сьогодні важко зібрати достатню кількість необхідної демографічної інформації через проблеми конфіденційності в Інтернеті, що обмежує використання даного підходу. Його все ще поєднують з іншими рекомендаціями як техніку підсилення для кращої якості.

Необхідно відзначити підхід побудови рекомендаційної системи за допомогоюд фільтрування на основі знань, що відповідає вимогам користувачів, і створювати рекомендації відповідно [42]. Особливим типом даного підходу є обмеження на основі рекомендаційної системи, які здатні

рекомендувати складні предмети, які рідко купуються (тобто автомобілі чи будинки), і виявляють важливі обмеження для користувача (ціна) [43].

Неможливо успішно використовувати колаборативну фільтрацію або фільтрацію на основі змісту у цьому домені елементів, оскільки доступно мало даних про взаємодію користувача з системою.

Багато дослідників почали комбінувати стратегії рекомендацій різними способами, створюючи гібридні рекомендаційні системи. Гібридні рекомендаційні системи об’єднують дві або більше інших стратегій з метою підсилення їхніх переваг і зменшення недоліків або обмежень.

Є кілька поширених типів комбінування:

– реалізація окремо колаборативних та контентних алгоритмів та поєднання їх припущень;

– включення деяких контентних правил до колаборативної методики;

– включення деяких колаборативних правил до контентної методики;

– побудова загальної моделі, що включає правила обох методик.

Однією із перших систем був мета-рівневий рекомендувач, який використовувався для пропонування веб-сайтів [44]. Він об’єднав CF для пошуку користувачів із подібними уподобаннями на веб-сайті та CBF для пошуку веб-сайтів із подібним вмістом. Невдовзі з’явилися інші роботи, такі як [45], і гібридні рекомендаційні системи стали загальновизнаним рекомендаційним підходом.

Розглянемо список найпоширеніших комбінацій технік, які утворюють гібридні рекомендаційні системи, а також проблеми, з якими найчастіше пов’язана кожна з цих комбінацій.

Прикладом такої комбінації є використання гібридної рекомендаційної системи [46], щоб покращити продуктивність багатокритеріальної рекомендації. Дана система визначає, що лише кілька критеріїв вибору впливають на вподобання користувачів щодо елементів та їхні відповідні оцінки. Кластеризація спочатку використовується для групування користувачів на основі критеріїв елементів, які вони віддають перевагу. Потім

рекомендаційна система використовується в кожному кластері подібних користувачів для прогнозування рейтингів. Інша спроба покращити прогностичну точність традиційної рекомендаційна системи представлено в роботі [47], де автори інтегрують у рекомендаційну систему окремі демографічні дані про користувачів, такі як стать, вік, професія тощо. Нечітка логіка використовується для обчислення подібності між користувачами, використовуючи ці додаткові демографічні дані, та інтеграції додаткових подібностей із подібностями на основі користувачів, розрахованими на основі рейтингів. історії. Після обчислення кінцевих подібностей користувачів їхній алгоритм прогнозує значення рейтингу. Додаткова продуктивність, отримана за рахунок кращої схожості користувачів, досягається ціною трохи більшого обчислювального часу, який прийнятний.

Ще однією популярною комбінацією гібридної рекомендаційної системи є система, яка придатна для роботи з великими обсягами даних і вирішення проблем масштабованості в системах електронної комерції [48]. Їхнє рішення спочатку включає колаборативну фільтрацію (Collaborative Filtering - CF) (коефіцієнти моменту добутку Пірсона) для зменшення набору даних шляхом пошуку найближчих сусідів кожного користувача, відкидання решти та скорочення набору даних. Після цього відстань до межі фільтрацією на основі вмісту (Content-Based Filtering -CBF) використовується для визначення межі прийняття рішень щодо товарів, придбаних цільовим користувачем. Останній крок поєднує оцінку CF (коефіцієнт кореляції між двома клієнтами) з оцінкою відстані до межі (відстань між межею прийняття рішення та кожним елементом) у зваженій лінійній формі. Автори повідомляють про покращену точність їх гібридної рекомендаційної системи, що працює в скороченому наборі даних, порівняно з іншими існуючими алгоритмами, які використовують повні набори даних.

В роботі [49] автори пропонують гібридний рекомендаційний метод CF-CBF, який базується на байєсівських мережах. Ця модель, яку вони побудували, використовує ймовірнісні міркування для обчислення розподілу

ймовірностей за очікуваним рейтингом. Вага кожної рекомендованої стратегії (CF і CBF) вибирається автоматично, адаптуючи модель до конкретних умов задачі (її можна застосовувати до різних доменів). Автори демонструють, що їх поєднання CF і CBF підвищує точність рекомендацій. Інші дослідження включають подібні математичні моделі або конструкції (наприклад, нечітку логіку), щоб поєднати CF і CBF і отримати продуктивність або інші переваги.

В роботі [50] пропонується CF і CBF поєднуються разом із демографічним фільтром (Demographic Filtering -DF) для створення рекомендацій для груп подібних профілів (користувачів). Такі рекомендації особливо корисні в соціальних мережах онлайн. Метою авторів є надання хороших рекомендацій у ситуаціях розрідженості даних. Перший CBF використовується для аналізу рейтингів і атрибутів елементів. Потім CF викликається як другий етап каскаду для створення групових рекомендацій. При цьому DF використовується для підсилення CF у випадках розріджених профілів (користувачі з невеликою кількістю рейтингів).

Доцільно розглянути два елементи рекомендаційної системи, де «Елемент-предмет» і «Користувач-користувач» є двома формами рекомендувачів CF, які відрізняються способом формування околиць. Деякі дослідження поєднують обидва з метою покращення загальної ефективності CF. Прикладом є робота [51], де автори представляють гібридну структуру рекомендацій, яку вони називають Collaborative Filtering Topic Model (CFTM), яка враховує як відгуки користувачів, так і оцінки щодо елементів певної теми (або домену) в електронній комерції. Перший етап, який знаходиться в автономному режимі, виконує аналіз настроїв у відгуках, щоб обчислити схожість користувача або елемента. Другий етап каскаду використовує «Елемент-предмет» і «Користувач-користувач» (перемикання) для прогнозування рейтингів. Автори оцінюють використання 6 наборів даних різних доменів від Amazon і повідомляють, що їхній гібридний підхід працює краще, ніж традиційний CF, особливо в ситуаціях розрідженості.

Були також дослідження, в яких CBF поєднувався з іншою технікою X, такою як кластеризація тощо. Одним із прикладів є робота [52], де автори описують і використовують цікаве поняття стилю життя користувача. Вони вибирають демографічну інформацію, дані про споживчі кредити та вподобання телевізійних програм як показники способу життя та підтверджують їхню значимість, виконавши статистичний аналіз 502 користувачів. Найважливіші атрибути стилю життя закодовані у двійковому коді та використовуються для формування сусідства та рейтингів кожного користувача за допомогою кореляції Пірсона. Отриману повну матрицю автори називають матрицею псевдо «Користувач – Елемент». Потім він використовується для прогнозування на основі Pearson (класичного CF) оригінальних оцінок елементів користувача. Повідомляється про значні покращення продуктивності.

Інші реалізації включають комбінації тієї самої стратегії рекомендацій, з різними мірами подібності або параметрами налаштування кожна, які використовуються в соціальних спільнотах, прототипи з використанням аналізу правил асоціації, нейронні мережі, генетичні алгоритми , зменшення розмірності, соціальні теги, семантичні онтології, аналіз шаблонів або різні класифікатори машинного навчання.

1.3 Аналіз інфокомунікаційних засобів для забезпечення підтримки процесів в рекомендаційних системах

Проведення аналізу інфокомунікаційних засобів для забезпечення підтримки процесів у рекомендаційних системах включає дослідження інструментів та технологій, що використовуються для збору, обробки та обміну інформацією у контексті рекомендацій. До таких засобів відносяться наступні:

1. Інфраструктура для збору даних, що включає системи та методи збору даних. Системи використовуються для отримання інформації про користувачів та об'єкти, такі як веб-аналітика, додатки для збору даних, IoT-

пристрої. Збір даних для рекомендаційних систем може використовувати різні технології для отримання відомостей про користувачів та об'єкти, які служать основою для створення рекомендацій. Серед інструментів потрібно виділити Web Scraping, який використовується для автоматичного збору даних з веб-сайтів, використовуючи скрипти, які аналізують HTML-код сторінок. Метод Web Scraping включає Python-бібліотека для парсингу HTML і XML, фреймворк Scrapy для створення веб-скраперів та витягування даних з веб-сайтів, а також Selenium для автоматизованого тестування веб-додатків, але також може використовуватися для скрапінгу даних, імітуючи браузерну поведінку.

Використання API (Application Programming Interface) надає зручний спосіб доступу до даних та функцій інших систем. Деякі API, які можуть бути корисними для рекомендаційних систем включають Social Media APIs (Twitter, Facebook, Instagram), як доступ до соціальних мереж для отримання даних про користувачів та їхніх взаємодій, Google Maps API для отримання геолокаційних даних та E-commerce APIs для отримання даних про товари та покупки користувачів.

Іншим інструментом є IoT Devices, який дозволяє здійснювати збір даних від пристроїв Інтернету речей, тобто IoT пристрої генерують велику кількість даних, які можна використовувати для створення персоналізованих рекомендацій. Інфраструктура IoT для збору даних та включають сенсори та датчики для збору даних про середовище, поведінку або здоров'я користувачів. Також важливими елементами є сережі зв'язку та хмарні платформи для передачі та збереження даних.

Збір даних від IoT пристроїв може допомогти у створенні контекстно-залежних рекомендацій, використовуючи інформацію про поведінку користувача в реальному часі.

Ці методи збору даних є важливими для створення рекомендаційних систем, оскільки дозволяють отримувати необхідні дані про користувачів та об'єкти для аналізу та генерації персоналізованих рекомендацій.

2. Засоби обробки та аналізу даних. Етап обробки та аналізу даних є одним із ключових при побудові рекомендаційної системи. Такі процеси допомагають зрозуміти поведінку користувачів та властивості об'єктів, щоб створити ефективні та персоналізовані рекомендації. До даних процесів відносять підготовку даних, а саме видалення дублікатів, обробка відсутніх значень та некоректних даних, приведення даних до одного масштабу для однорідності у порівнянні, а також вибір найбільш інформативних ознак для покращення ефективності алгоритмів.

Аналіз даних включає процес розуміння основних характеристик даних через використання статистичних метрик, візуалізацію даних через графіки та діаграми для відображення взаємозв'язків та патернів у даних, а також кластеризацію через групування схожих об'єктів або користувачів для подальшої рекомендації в межах одного кластера.

Процес моделювання даних відбувається через коллаборативний та контентний підходи, а саме використання моделей, які враховують спільну поведінку користувачів або властивості об'єктів Використання алгоритмів для рекомендаційних систем, таких як нейронні мережі, регресія, класифікація, ансамблеві методи тощо. Визначення ефективності різних моделей відбувається за допомогою метрик, таких як точність, відновлення та F1-оцінка.

Ще одним етапом обробки та аналізу даних є персоналізація та оновлення рекомендацій, а саме використання інформації про користувачів для створення унікальних рекомендацій, а також адаптація до змін, що відбувається через системи, які змінюють рекомендації на основі нових даних або зміни у поведінці користувачів.

Дані процеси дозволяють рекомендаційним системам аналізувати та використовувати великі обсяги даних для створення точних та значущих рекомендацій, що відповідають індивідуальним потребам користувачів.

3. Засоби комунікації та взаємодії користувачів з системою та подання рекомендацій. Взаємодії в рекомендаційних системах реалізуються через

подання рекомендацій, що враховують індивідуальні вподобання користувачів та доступ через веб-інтерфейс, мобільні додатки, email, push-сповіщення тощо. Важливим елементом виступає взаємодія з користувачем, що включає збір фідбеку через запит на відгук користувача щодо рекомендацій для поліпшення системи та особисті налаштування, тобто можливість користувача вибирати параметри рекомендацій.

Персоналізація комунікації в рекомендаційних системах реалізується через адаптовані повідомлення, а саме врахування контексту та інформації про користувача для створення індивідуальних сповіщень та часовий та контекстуальний аналіз, а саме подання рекомендацій відповідно до актуального часу або поточних подій.

Ефективність взаємодії в даних системах включає тестування та аналіз взаємодії, тобто визначення та вдосконалення каналів комунікації, які забезпечують найкращі результати та оптимізація UX/UI, а саме створення інтерфейсів, що сприяють зручній взаємодії та розумінню рекомендацій.

Всі процеси сприяють покращенню досвіду користувача та ефективності рекомендаційної системи, дозволяючи взаємодіяти з користувачами у зручний та персоналізований спосіб.

4. Засоби захисту даних та забезпечення приватності. Ці питання є критичними аспектами у рекомендаційних системах, оскільки ці системи працюють із особистою інформацією користувачів. При створенні рекомендаційних систем слід враховувати безпеку даних, що включає захист конфіденційності даних користувачів та реалізується через шифрування для захисту передачі та збереження конфіденційної інформації та контроль доступу до даних, обмеження прав користувачів на рівні системи, а також моніторинг безпеки через системи виявлення вторгнень та моніторинг активності для вчасного реагування на загрози безпеці. Дотримання правил та регулювань відбувається відповідно до захисту персональних даних та забезпечення чіткого розуміння користувачем того, як їхні дані збираються та використовуються для рекомендацій, а також надання користувачам

можливості видалення, редагування чи контролю над своєю особистою інформацією.

Забезпечення безпеки та приватності даних у рекомендаційних системах вимагає дотримання високих стандартів безпеки, врахування вимог законодавства та створення довіри користувачів до обробки їхніх особистих даних.

5. Засоби оптимізації та моніторингу. У рекомендаційних системах оптимізація та моніторинг – це постійний процес покращення ефективності та точності рекомендацій через аналіз та вдосконалення системи, який включає аналіз даних та впровадження нових методів через тестування різних моделей та алгоритмів для вдосконалення точності рекомендацій, а також налаштування параметрів моделей для досягнення кращих результатів.

Даний процес включає також A/B тестування та експерименти, а саме порівняння двох версій системи для визначення кращої ефективності та експерименти для вибору оптимальної стратегії рекомендацій.

Моніторинг та відслідковування реалізується постійним спостереженням та аналізом з визначенням метрик та показників успішності для визначення та моніторингу ключових показників ефективності системи та вчасного виявлення змін у поведінці користувачів або результатів системи.

Підтримка та постійне вдосконалення системи відбувається на основі отриманих результатів та змін у вимогах користувачів, що дозволяє реагувати на запити та зворотний зв'язок користувачів для поліпшення сервісу. Дані процеси допомагають забезпечити, що рекомендаційна система залишається ефективною, точною та відповідає потребам користувачів у змінному середовищі. Постійна оптимізація та моніторинг є ключем до успіху рекомендаційних систем.

Проведений аналіз є важливим для розвитку рекомендаційних систем, оскільки він допомагає визначити оптимальні інструменти та технології для підтримки процесів збору, обробки, взаємодії та оптимізації в рекомендаційних системах.

1.4 Постановка наукового завдання та визначення мети дослідження

З урахуванням нагальних потреб підвищення якості процесів в сфері освіти, які стосуються автоматизованої підтримки прийняття рішень при вирішенні різноманітних завдань, виходячи з оцінки існуючих моделей, методів та підходів до побудови рекомендаційних систем, можна зробити висновок про те, що на сьогоднішній день актуальною задачею є розробка методики побудови рекомендаційної системи для сфери освіти на основі гібридного підходу.

Метою дослідження є підвищення якості надання рекомендацій в сфері освіти за рахунок використання рекомендаційної системи на основі гібридного підходу з використанням нечіткої логіки та методів обробки природньої мови.

Для розробки методики побудови рекомендаційної системи в сфері освіти на основі гібридного підходу та досягнення мети дисертаційного дослідження необхідно вирішити такі наукові завдання:

1. Провести аналіз моделей та методів обробки даних рекомендаційних систем.
2. Визначити характеристики рекомендаційної системи для сфери освіти.
3. Розробити модель представлення даних рекомендаційної системи для освітньої сфери на основі нечіткої логіки.
4. Провести моделювання структури рекомендаційної системи для сфери освіти.
5. Розробити метод формування рекомендації по об’єктам інтересу освітньої сфери на основі гібридного підходу.
6. Розробити алгоритм застосування методики побудови рекомендаційної системи для сфери освіти на основі гібридного підходу.
7. Визначити параметри моделей та методів рекомендаційної системи, розробити структури баз даних, провести моделювання архітектури системи

та розробити програмні засоби для задачі формування рекомендації в сфері освіти.

8. Провести дослідження результатів застосування розробленої методики побудови рекомендаційної системи в сфері освіти на основі гібридного підходу.

Висновки до розділу 1

Аналіз існуючих підходів до формування автоматизованих та автоматичних рекомендацій засобами рекомендаційних систем продемонстрував доцільність та актуальність їх використання для задач в сфері освіти.

В першому розділі розглянуто загальну постановку задач рекомендаційної системи, визначено перелік та вміст етапів по створенню та впровадженню рекомендаційних систем.

Проведено аналіз підходів до побудови рекомендаційних систем, зокрема, досліджено особливості використання колаборативної фільтрації, контентного підходу та гібридні підходи. Визначено переваги та недоліки кожного з методів та виконано огляд засобів для усунення типових недоліків рекомендаційних систем, побудованих з використанням того чи іншого підходу. Дослідження показало, що на поточний час актуальним є створення саме гібридних рекомендаційних систем, при чому структура та методи обробки даних такої системи повинні обов’язково враховувати контекст предметної галузі та кінцеві цілі рекомендації.

Проведено аналіз інфокомунікаційних засобів для забезпечення підтримки процесів в рекомендаційних системах. Визначено ключові інструменти та технології збору, обробки та обміну інформацією в контексті рекомендацій, розглянуто питання захисту конфіденційної інформації, підтримки та моніторингу якості надання рекомендацій.

Визначено мету дослідження та поставлено наукові завдання.

РОЗДІЛ 2.
АНАЛІЗ МОДЕЛЕЙ ТА МЕТОДІВ ОБРОБКИ ДАНИХ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ

2.1 Моделі та методи обробки експертних даних для задачі колаборативної фільтрації

Будь-який об’єкт, з метою оцінювання його властивостей, описують певною множиною показників, які відрізняються за змістом, значенням або якістю. Показники поділяють на дві категорії: якісні та кількісні. Кількісні показники – це величини, що пов’язані із фізичною сутністю вимірюваних властивостей об’єкту та дозволяють оцінювати його у конкретних одиницях вимірювання: абсолютних або відносних (відсотки, частини). Якісні показники – це показники, що не підлягають безпосередньому вимірюванню, а характеризують ступінь відповідності властивості об’єкту певним вимогам. Загальна оцінка об’єкту може бути визначена методами експертного опитування. Для цього усі показники властивостей повинні мати кількісну форму, тому якісні показники оцінюють певною кількістю балів відповідно до шкали оцінювання. Шкала оцінювання – це визначена система правил, за якими надається оцінка властивостям об’єктів. Кожна шкала має характеристики: опис, порядок, відстань, точка відліку [53, 54, 55].

Використовують наступні типи шкал:

1. Номінальна шкала (шкала найменувань). Має лише одну характеристику «опис», характеризується оцінкою еквівалентності до вимірюваної властивості та підтримує операцію порівняння. Оцінка властивості об’єкту за номінальною шкалою виконується порівнянням її з еталоном у формі «дорівнює» чи «не дорівнює» та відповідає значенню за бінарною шкалою 0 або 1. Прикладом використання номінальної шкали в освітній сфері можуть бути результати перевірки викладачем проходження студентами закладу освіти випробування за певною навчальною дисципліною,

розділом, або темою по формі заліку. Позначкою «залік» оцінюється успішне проходження, позначкою «незалік» не проходження випробування.

2. Порядкова (рангова) шкала [56]. До шкали додається характеристика «порядок», що передбачає визначення ступеню вираження вимірюваної властивості та дозволяє упорядковувати об’єкти операціями «більше» та «менше». В результаті об’єкт визнається кращім (гіршим) від іншого за певним критерієм. Прикладом використання рангової шкали в освітній сфері можуть бути результати участі студентів у конкурсі студентських робіт. Члени конкурсної комісії визнають, яка робота є кращою за інші та отримує перше місце, яка гірша за першу, але найкраща за інші, та отримує друге місце і так далі.

3. Інтервальна (метрична) шкала. До шкали додається характеристика «відстань», що передбачає визначення не тільки відмінності ступеню вираження вимірюваної властивості у об’єктів, а й наскільки ці ступені відмінні – відстань між ними. Використовуючи операції «додавання» та «віднімання» шкала дозволяє не тільки визначати чи є об’єкт кращім (гіршим) від іншого, але й наскільки кращим (гіршим) [57, 58, 59]. Прикладом використання інтервальної шкали в освітній сфері можуть бути результати перевірки викладачем проходження студентами закладу освіти випробування за певною навчальною дисципліною у формі іспиту. Результати перевірки проводяться за шкалою з наступних позначок: «A», «B», «C», «D», «E», «FX» та «F», при цьому кожній оцінці відповідає числовий інтервал, наприклад: «A» - [90; 100], «B» - [82; 89], «C» - [75; 81], «D» - [65; 74], «E» - [60-64], «FX» - [35-59], «F» - [0; 34].

4. Шкала відношення (абсолютна шкала). До шкали додається остання характеристика «точка відліку», тобто істинна нульова точка, що відповідає повній відсутності вимірюваної властивості у об’єкта. Використовуючи операції «множення» та «ділення» шкала дозволяє визначити не тільки наскільки є об’єкт кращім (гіршим) від іншого, а й в скільки раз [60]. Таким чином, ця шкала підтримує усі математичні операції та може бути використана

для тих властивостей, які можливо виразити кількісно. Прикладом використання шкали відношення в освітній сфері може бути оцінка кількості студентів в аудиторії. Нульова точка шкали буде відповідати відсутності студентів в аудиторії. Кількість студентів у певний день може бути вдвічі більшою (меншою) за кількість у інший день [61].

В залежності від природи об’єктів для оцінки їх вимірюваних властивостей можуть використовуватися різні шкали.

1. Ранжування – методика оцінки об’єктів за порядковою шкалою шляхом розташування об’єктів за зростанням (зменшенням) ступеню вираження вимірюваної властивості. Задача ранжування об’єктів за ступеню вираження властивостей є однією з найпоширеніших задач експертного оцінювання. Визначають також задачу суворого ранжування об’єктів, у який ступень вираження властивостей декількох об’єктів не може бути однаковою. При ранжуванні використовують порядкову шкалу оцінювання, а у результаті ранжування отримують ранжований ряд об’єктів [62, 63].

З метою підвищення об’єктивності, ранжування об’єктів проводиться групою експертів [64, 65], тоді кожен об’єкт повинен отримати оцінку від кожного експерту. Ранг, що приписується об’єкту експертом можна визначити як x_{ij}, де $i = \overline{1,n}$ – певна вимірювана властивість об’єкту, n – кількість властивостей, $j = \overline{1,m}$ – певний експерт, m – кількість експертів. При цьому діапазон значень рангу об’єкту $x_{ij} \in [1; N]$, де N – кількість об’єктів. Значення $x_{ij} = 1$ відповідає найкращому об’єкту, $x_{ij} = N$ – найгіршому за вимірюваної властивості. Результати ранжування дозволяють обрати найсуттєвіші властивості з усієї сукупності властивостей об’єкту.

Отримані від експертів оцінки надаються у вигляді таблиці 2.1, яка у подальшому обробляється методами математичної статистики для отримання узагальненої експертної оцінки [66].

У якості переваги методу зазначають його зрозумілість та простоту. До недоліків методу відносять наступні:

– складність при кількості властивостей об’єктів, що перевищує 15-20, бо це вимагає врахування складних з’язків між ранжованими об’єктами;

– складність ранжування у випадку дуже малої різниці ступеню вираженості вимірюваної властивості між ними;

– не є зрозумілим наскільки далеко один від одного знаходяться об’єкти за вимірюваною властивістю [56].

Таблиця 2.1 – Загальний вигляд матриці ранжування

Експерти / Властивості	1	2	…	j	…	m
1	x_{11}	x_{12}	…	x_{1j}	…	x_{1m}
2	x_{21}	x_{22}	…	x_{2j}	…	x_{ij}
…	…	…	…	…	…	…
i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{im}
…	…	…	…	…	…	…
n	x_{n1}	x_{n2}	…	x_{nj}	…	x_{nm}

2. Попарне порівняння – методика оцінки об’єктів в усіх можливих парах, визначення, який з об’єктів у кожній парі кращій, гірший чи дорівнює другому за вимірюваної властивості [67, 68]. Прикладом використання попарного порівняння можна навести метод аналізу ієрархій (Analytical Hierarchy Process) – розроблена американським математиком Томасом Л. Сааті структурована техніка для організації та аналізу складних рішень за безлічі критеріїв [69, 70]. Головна ідея методу – це порівняння альтернативних рішень за багатьма критеріями шляхом попарного їх порівняння. При попарному порівнянні використовується порівняльна шкала оцінювання. Методики попарного порівняння характеризуються високим рівням достовірності отриманих оцінок, завдяки чому отримали найбільш широке розповсюдження.

При виконання попарного порівняння експерти розглядають множину пар вимірюваної властивості об’єктів та для кожної пари встановлюють який саме об’єкт є більш значущий за вимірюваної властивості, або приходять до висновку про їх еквівалентність. Значення a_{ij}, що приписується об’єктам *i* та *j*, може бути визначена наступним чином (2.1):

$$a_{ij} = \begin{cases} 2, \text{якщо об'єкт } i \text{ більший за об'єкт } j \ (i > j); \\ 1, \text{якщо об'єкт } i \text{ дорівнює об'єкту } j \ (i = j); \\ 0, \text{якщо об'єкт } i \text{ меньший за об'єкт } j \ (i < j). \end{cases} \quad (2.1)$$

Надана схема заповнення матриці не є єдиною можливою, відповідно до природи об'єкту і задачі, що вирішується, можуть біти використані і інші види схем.

Отримані від експертів значення a_{ij} надаються у вигляді таблиці 2.2.

Таблиця 2.2 – Загальний вигляд матриці попарного порівняння

Властивості	1	2	…	j	…	n	∑
1		a_{12}	…	a_{1j}	…	a_{1n}	$\sum_{j=1}^{n} a_{1j}$
2	a_{21}		…	a_{2j}	…	a_{2j}	$\sum_{j=1}^{n} a_{2j}$
…	…	…		…	…	…	
i	a_{i1}	a_{i2}	…		…	a_{in}	$\sum_{j=1}^{n} a_{ij}$
…	…	…	…	…		…	
n	a_{n1}	a_{n2}	…	a_{nj}	…		$\sum_{j=1}^{n} a_{nj}$

Останній стовпець таблиці заповнюється сумою значень за відповідним рядком та показує відносну вагу об'єкту за відповідною властивістю. Об'єкт, що за результатами оцінювання буде мати найбільше значення, буде вважатися найкращім за вимірювані властивості.

Сумарні оцінки властивостей, що надаються багатьма експертами, отримують побудовою сумарної матриці зі значеннями $\overline{a_{ij}}$, що є сумами значень порівняння ознак, які надані всіма експертами. Сумарна матриця матриця має вигляд, наведений у таблиці 2.3.

Таблиця 2.3 – Загальний вигляд сумарної матриці попарного порівняння

Властивості	1	2	…	j	…	n	∑
1		$\sum_{k=1}^{m} a_{1j}^{(k)}$	…	$\sum_{k=1}^{m} a_{ij}^{(k)}$	…	$\sum_{k=1}^{m} a_{1n}^{(k)}$	$\sum_{k=1}^{m}\sum_{j=1}^{n} a_{1j}^{(k)}$
2	$\sum_{k=1}^{m} a_{21}^{(k)}$		…	$\sum_{k=1}^{m} a_{2j}^{(k)}$	…	$\sum_{k=1}^{m} a_{2j}^{(k)}$	$\sum_{k=1}^{m}\sum_{j=1}^{n} a_{2j}^{(k)}$
…	…	…		…	…	…	
i	$\sum_{k=1}^{m} a_{i1}^{(k)}$	$\sum_{k=1}^{m} a_{12}^{(k)}$	…		…	$\sum_{k=1}^{m} a_{in}^{(k)}$	$\sum_{k=1}^{m}\sum_{j=1}^{n} a_{ij}^{(k)}$
…	…	…	…	…		…	
n	$\sum_{k=1}^{m} a_{n1}^{(k)}$	$\sum_{k=1}^{m} a_{n2}^{(k)}$	…	$\sum_{k=1}^{m} a_{nj}^{(k)}$	…		$\sum_{k=1}^{m}\sum_{j=1}^{n} a_{nj}^{(k)}$

До переваг методу попарного порівняння відносять можливість оцінки великої кількості об’єктів та можливість оцінки у випадку незначної різниці ступеню вираження вимірюваної властивості між об’єктами, що є складними при виконанні оцінки ранжуванням. Прикладом використання попарної оцінки в освітній сфері може бути задача порівняння дисциплін вільного вибору студента – кожну пару дисциплін студент може порівнювати між собою для того, щоб визначити яка з низ має перевагу над іншою. Також таку систему можна використовувати для порівняння керівників диплому, літературних джерел та інших об’єктів інтересу користувачів.

3. Безпосередня оцінка (бальний метод) – методика оцінки об’єктів, що дозволяє не тільки ранжувати об’єкти, але й визначати, на скільки одна вимірювальна властивість є більш вагомою за інші.

При виконанні безпосередньої оцінки використовується інтервальна шкала оцінювання. Процедура оцінювання передбачає побудову шкали інтервалів розбиттям діапазону зміни вимірювальної властивості об’єкта на окремі інтервали, кожному з яких приписуються певні числові значення, що є оцінкою, або балом. В процесі оцінювання експерти розміщують об’єкт на

шкалі інтервалів відповідно до ступеня вимірювальної властивості, тобто ставлять у відповідність об’єкту інтервал с певною кількістю балів [71].

Кількість інтервалів, на які розбивається діапазон зміни властивості, може бути довільною, може відрізнятися для різних експертів, окрім того, методикою дозволяється встановлення декільком об’єктам однакового значення балів. З метою посилення умов та збільшення точності вимірювання діапазон зміни властивості може бути представлено безперервною числовою віссю. Також методика безпосередньої оцінки не висуває жорстких умов, що до використання саме числової вісі. Наприклад, таку властивість, як колір об’єкту, експертам може бути незрозуміло представляти у вигляді чисел, числових кодів, або певного значення спектру[72].

Прикладом безпосередньої оцінки в можуть бути результати перевірки викладачем рівня виконання курсової роботи студентів закладу освіти. Інтервал, що характеризує виконану роботу розбивається на 5 інтервалів. Найвища кількість балів 5 надається студентові, робота якого виконана майже без помилок, найнижча 2 бали за неприйнятний рівень виконаної роботи (див.рис.2.1).

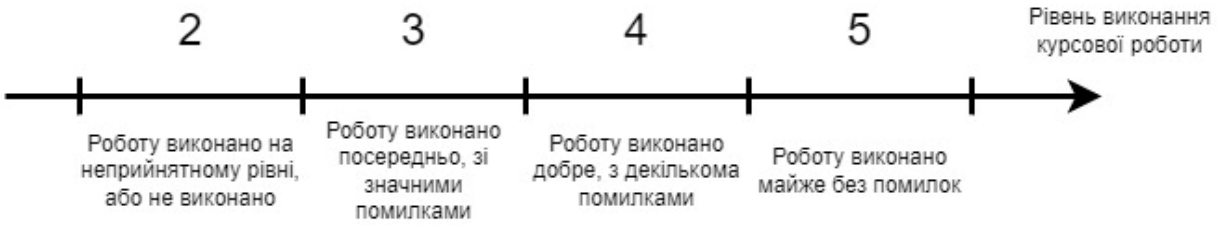


Рис. 2.1. Приклад розбиття діапазону вимірюваної властивості об’єкта на інтервали

Оцінка об’єкту не обмежується лише їх ранжуванням або порівнянням. Не менш важливими аспектами є узагальнена думка експертів та рівень узгодженості думок що до оцінки властивостей об’єкту.

Для отримання узагальненої думки експертів, що до об’єкту який досліджується, індивідуальні оцінки експертів обробляють з використанням якісних та кількісних методів і найчастіше для цього використовуються методи математичної статистики.

1. Узагальнена експертна оцінка. Якщо склад групи експертів є однорідним, то для формування узагальненою експертної оцінки рангу об’єкту за вимірювальної властивості використовують такі статистичні величини, як середнє арифметичне та медіана рангів об’єктів [72].

Середнє арифметичне рангів об’єктів $\overline{x_i}$ за вимірювальної властивості i розраховують за формулою (2.2):

$$\overline{x_i} = \frac{\sum_{j=1}^{m} x_{ij}}{m}, \tag{2.2}$$

де x_{ij} – оцінка i-властивості, що надана j-м експертом, $i = \overline{1,n}$, $j = \overline{1,m}$; m – кількість експертів; n – кількість властивостей.

За результатами розрахунків середніх рангів будують підсумковий ранжований ряд для вимірювальної властивості i. Об’єкт, що має найменше значення середнього рангу відповідає найкращому об’єкту, а об’єкт, що має найбільше значення n – найгіршому за вимірюваної властивості i.

Якщо склад групи експертів не є однорідним, то серед групи експертів визначають фахівців, які вважаються провідними та чиї оцінки повинні бути більш значимими. У цьому випадку кожному експерту встановлюється певний ваговий коефіцієнт – K_j, який враховується при обчисленні середнього рангів наступним чином (2.3):

$$\overline{x_i} = \frac{\sum_{j=1}^{m} K_j \cdot x_{ij}}{m}, \tag{2.3}$$

де K_j – ваговий коефіцієнт j-го експерта.

З точки зору освітньої сфери такими експертами можуть виступати, наприклад, фахівці-практики, що надають рекомендацію щодо корисності та

доцільності вивчення певної дисципліни або якості тих чи інших навчально-методичних матеріалів; випускники закладу, які оцінюють роботу викладачів чи якість тих чи інших дисциплін; рецензенти методичних матеріалів, посібників, підручників тощо.

Медіана рангів обчислюється наступною послідовністю кроків:

1) упорядкування даних: ранги об'єкту за вимірювальної властивості сортуються за зростанням;

2) обчислення медіани: у упорядкованій послідовності рангів знаходять медіану – значення рангу, що поділяє послідовність таким чином, що кількість рангів з більшим або рівним значенням до обраної медіани дорівнюють кількості рангів з меншим або рівним значенням;

3) ранжування: за обчисленими медіанами рангів будують ітоговий ранжований ряд, де об'єкт, з найменшим значенням медіани рангу відповідає найкращому об'єкту, а об'єкт, з найбільшим значенням медіани рангу – найгіршому за вимірюваної властивості *i*.

Об'єкти, що досліджуються, вимірюються множиною властивостей, що мають різний вплив на оцінку об'єкту. Для визначення, яка з властивостей є більш впливовою, обраховують відносну вагу властивостей, що є важливим показником для побудови персональних рекомендацій [73].

Відносна вага властивостей для кожного об'єкту визначається наступним чином. Нехай x_{ij} – оцінка *i*-властивості що надана *j*-м експертом, тоді вагу w_{ij} властивості *i* за оцінкою *j*-го експерта визначають як (2.4):

$$w_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}}, \tag{2.4}$$

де w_{ij} – вага властивості i за оцінкою *j*-го експерта; x_{ij} – оцінка *i*-тої властивості, що надана *j*-м експертом, $i = \overline{1,n}$, $j = \overline{1,m}$; m – кількість експертів; n – кількість властивостей.

Тоді відносну вагу w_i властивості i об'єкту за оцінкою усіх експертів, можна визначити наступним чином (2.5):

$$w_i = \frac{\sum_{j=1}^{m} w_{ij}}{m}, \tag{2.5}$$

де w_i – відносна вага i-властивості об’єкту за оцінками усіх експертів; w_{ij} – вага властивості i за оцінкою j-го експерта; m – кількість експертів; n – кількість властивостей.

Властивість з більшою відносною вагою вважається більш вагомою для даного об’єкту.

2. Ступень узгодженості експертних оцінок. Якщо оцінювання об’єкту виконується групою експертів, результати оцінювання будуть мати певні розбіжності. Така оцінка вважається надійною тільки тоді, коли оцінки різних експертів добре узгоджені. Для аналізу узгодженостей експертних оцінок використовують такі статистичні величини, як середнє квадратичне відхилення, коефіцієнт варіації та розмах варіації [72, 74].

Середньоквадратичне відхилення допомагає визначити на скільки в середньому відхиляються конкретні значення оцінок експертів від їх середнього значення, обчислення якого виконується за формулою (2.2). Обчислення середньоквадратичне відхилення експертних оцінок для i-ї властивості об’єкту виконують за (2.6):

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{m}(x_{ij} - \bar{x}_i)^2}{m-1}}, \tag{2.6}$$

де σ_i – середньоквадратичне відхилення експертних оцінок i-ї властивості, $i = \overline{1,n}$; x_{ij} – оцінка i-властивості, що надана j-м експертом, $i = \overline{1,n}$, $j = \overline{1,m}$; $\bar{x}_i$ – середнє арифметичне оцінок властивості i групою експертів; m – кількість експертів; n – кількість властивостей.

Коефіцієнт варіації дозволить визначити співвідношення середньоквадратичного відхилення та середньої величини значення оцінок експертів. Виражається коефіцієнт варіації у відсотках та визначається (2.7):

$$V_i = \frac{\sigma_i}{\bar{x}_i} \cdot 100\%, \tag{2.7}$$

де V_i – коефіцієнт варіації оцінок i-ї властивості; σ_i – середньоквадратичне відхилення експертних оцінок i-ї властивості, $i = \overline{1,n}$, обчислюється за формулою (2.6); $\overline{x_i}$– середнє арифметичне оцінок властивості i групою експертів; n – кількість властивостей.

Розмах варіацій визначає різницю між найбільшим і найменшим значеннями оцінки властивості. Обчислюється розмах варіації для оцінок i-ї властивості об'єкту наданих групою експертів за формулою (2.8):

$$R_i = max_i - min_i, \tag{2.8}$$

де R_i – розмах варіації оцінок i-ї властивості об'єкту; $max_i = max_{j=1}^{m} x_{ij}$ – максимальне значення оцінки i-ї властивості об'єкту, що надана групою експертів; $min_i = mix_{j=1}^{m} x_{ij}$ – мінімальне значення i-ї властивості об'єкту, що надана групою експертів [75].

При оцінюванні властивості об'єкту методом ранжування, коли оцінки надаються не числовими значеннями, а у вигляді рангів, узгодженість експертних оцінок визначають частіш за все за допомогою коефіцієнтів кореляції та коркондації, наприклад, коефіцієнт рангової кореляції Спірмена, коефіцієнт рангової коркондації Кендалла [76, 77].

Коефіцієнт рангової кореляції Спірмена дозволяє оцінити узгодженість між двома ранжованими рядами. Коефіцієнт приймає значення у діапазоні [-1;1]: значення -1 показує зворотню залежність між рангами, що свідчить про велике розходження у думках експертів, 0 – відсутність залежності, 1 – пряму залежність між рангами ознак, що свідчить про повну узгодженість думок експертів.

Узгодженість між ранжованими рядами вимірюваної властивості усіх об'єктів, що побудовано двома експертами обчислюють за формулою (2.9):

$$\rho = 1 - \frac{\sum_{i=1}^{n}\left(x_{ij} - x_{ik}\right)^2}{\frac{1}{6}(n^3 - n)}, \tag{2.9}$$

де ρ – коефіцієнт рангової кореляції Спірмена між ранжованими рядами вимірюваної властивості усіх об'єктів двох експертів; x_{ij} – ранг, що надано i-му об'єкту j-м експертом за вимірюваної властивості; x_{ik} – ранг, що надано i-му об'єкту k-м експертом за вимірюваної властивості; n – кількість об'єктів.

Коеффіцієнт конкордації Кендалла (множинний коефіцієнт рангової кореляції) дозволяє оцінити узгодженість більш ніж між двома ранжованими рядами. Коефіцієнт приймає значення у діапазоні [0;1]: значення 0 показує повну відсутність узгодженості між оцінками експертів, 1 – повну узгодженість думок експертів [78].

Узгодженість між ранжованими рядами вимірюваної властивості усіх об'єктів, що побудовано m експертами обчислюють за допомогою (2.10):

$$W=\frac{12}{m^2(n^3-n)}\sum_{i=1}^{n}\left(\sum_{j=1}^{m} x_{ij} - \frac{1}{2}m(n-1)\right)^2, \qquad (2.10)$$

де W– коефіцієнт конкордації між ранжованими рядами вимірюваної властивості усіх об'єктів для групи експертів; x_{ij} – ранг, що надано i-му об'єкту j-м експертом за вимірюваної властивості; n – кількість об'єктів; m – кількість експертів.

2.2 Моделі та методи обробки даних рекомендаційної системи на основі нечіткої логіки

Класична логіка не здатна оперувати нечітко визначеними поняттями, оскільки усі формальні системи оперують лише з висловлюваннями, що мають два визначених стани, як то «істина», або «не істина», які позначаються значеннями «1» та «0» відповідно. Теорія нечітких множин забезпечує іншу парадигму та міркує у термінах невизначеності. Лотфі А. Заде запропонував теорію нечітких множин і пов'язану з нею логіку, а саме нечітку логіку [79]. По суті, нечітка множина – це множина, члени якої можуть мати ступені приналежності між 0 і 1, на відміну від класичних множин, де кожен елемент

повинен мати або 0, або 1 як ступінь приналежності. Оскільки класична логіка базується на класичній теорії множин, нечітка логіка базується на нечіткій множині.

Розглянемо деяку множину X та її підмножину A: $A \subset X$, елемент x є елементом множин X та A, $x \in X$ та $x \in A$. Належність x до A можна описати характеристичною функцією $\mu_A(x)$:

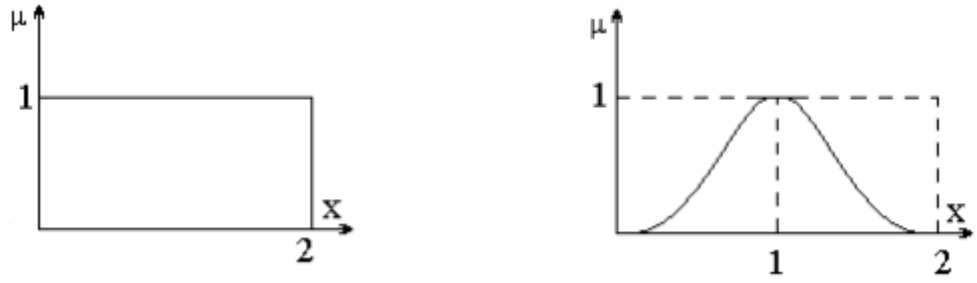
$$\mu_A(x) = \begin{cases} 1, x \in A; \\ 0, x \notin A. \end{cases} (2.11) . \qquad (2.11)$$

Елементи множини X або належать, або не належать множині A, характеристична функція приймає значення 0 або 1 відповідно. Припустимо, що функція може приймати будь яке значення з інтервалу [0;1]. Наприклад:

1) x не належить A, тоді $\mu_A(x)$ = 0;
2) x належить до A незначною мірою, тоді значення $\mu_A(x)$ близьке до 0;
3) x більш-меньш належить до A, тоді значення $\mu_A(x)$ однаково близьке як до 0, так і до 1;
4) x належить до A значною мірою, тоді значення $\mu_A(x)$ близьке до 1;
5) x не належить A, тоді $\mu_A(x)$ = 1.

Таким чином, можемо описати поняття нечіткої множини. Розглянемо множину X, нечітку підмножину A в X визначимо як безліч пар $(x, \mu_A(x))$, де $x \in X$, а функція $\mu_A(x): X \longrightarrow [0; 1]$. Функція $\mu_A(x)$ називається функцією належності нечіткої підмножини A та вказує ступінь належності елемента x до множини A. Ступінь належності $\mu_A(x)$ є суб'єктивною мірою того, наскільки елемент $x \in X$ відповідає нечіткій множині A [80].

На рис. 2.2 наведено графіки функції належності для звичайної множини чисел $B = \{x | 0 \leq x \leq 2\}$ та нечіткої підмножини $C = \{x | "x \text{ близьке до } 1"\}$, при чому вигляд функції належності μ_C не є чітко визначеним, а залежить від поняття «близький» у конкретній ситуації.



а) б)

Рис. 2.2. Графіки функцій належності

(а – звичайна множина, б – нечітка підмножина)

Функції належності можуть задаватися явним переліком елементів та відповідним до них конкретним значенням функції належності, або аналітично у вигляді формул [81]. Функції належності класифікують за виглядом та формою їх графічного подання. Найпоширенішими у використанні на практиці є наступні:

– кусково-лінійні функції належності: трикутна, трапецієподібна, лінійно Z-образна, лінійно S-образна;

– квадратична;

– експотенціальна (гаусівська):

– сигмоїдальна (логістична);

– дзвоноподібна.

Для побудови функцій належності використовують прямі та непрямі методи [82, 83, 84, 85].

Прямі методи передбачають явний опис функції належності, завдання експертом значення $\mu(x)$ для кожного $x \in X$. Такі методи використовуються до властивостей об’єктів, що можуть бути виміряні – якісні показники. Для об’єкта виділяють набір властивостей та визначають для кожної з них полярні значення, що будуть відповідати граничним значенням функції належності 0 та 1. Використовуючи приведену шкалу, експерт задає $\mu_A(x) \in [0; 1]$ та формує векторну функцію належності $\{\mu_A(x_1), \mu_A(x_2), \dots, \mu_A(x_n)\}$. Одним з різновидів методів є прямі групові методи, коли группа експертів визначає

належність, або неналежність, об’єкту до певної множини. Тоді значення функції належності об’єкту визначається відношенням позитивних відповідей експертів до загального числа експертів у групі.

Непрямі методи передбачають використання правил, або експертних знань, використовуються до властивостей об’єктів, що не можуть бути виміряні чисельно-якісними показниками. Прикладом може бути використання методу парних порівнянь. Експерт формує матрицю порівнянь M (табл. 2.2), а побудова вектору виконується пошуком вектору $M_a = \lambda_{max} a$, де λ_{max} – найбільше власне значення матриці M.

При оцінюванні експертами властивостей об’єктів на практиці існують різні ступені невизначеності та нечіткості. Для врахування цих аспекти та отримання достовірних оцінок від експертів використовують нечітку логіку. Оцінки експертів при використанні нечіткої логіки можуть бути представлені наступними способами:

1) функції належності – оцінка експерта описується функцією належності та вказує на ступінь належності конкретного значення до наданої експертом оцінки;

2) лінгвістична змінна – оцінку можна надати лінгвістичною змінною, що є характеристикою оцінки;

3) множина термів – оцінки можуть бути надані експертом певними термінами з множини термів лінгвістичної змінної;

4) висновок на основі правил – оцінка експерта може бути отримана з використанням правил нечіткої логіки на основі нечітких оцінок;

5) нечіткі числа – замість точних чисел, для представлення конкретної оцінки використовують числа, що не мають однозначних границь (нечіткі числа).

Розглянемо метод визначення оцінок експертів за допомогою функцій належності.

Розглянемо задачу оцінки експертами певного об’єкту: група з m експертів повинна дати числову оцінку властивостям об’єкту, треба визначити

колективну оцінку об’єкту. Відповідно до «чіткої» постановки задачі експерти повинні дати оцінку об’єкту у вигляді числового вектору, але у реальних умовах експерти частіше за конкретні числові значення оперують висловами, наприклад, «оцінка близька до числа b», або «оцінка лежить у межах [b,c]». Таким чином, доцільно поставити нечітку задачу числової оцінки об’єкту, де експертам треба надати оцінку, використовуючи певний перелік тверджень, наприклад:

1) «оцінка лежить у межах [b,c]»;
2) «оцінка близька до b»;
3) «оцінка не більша за b»;
4) «оцінка не менша за b».

Тоді, оцінки експертів можна формалізувати та описати функціями належності.

Твердження «оцінка лежить у межах [b,c]» описується нечіткою підмножиною A_1 із трапецієподібною функцією належності (2.12) $\mu_{A_1}(x, a, b, c, d)$, графік якої наведено на рисунку 2.3.

$$\mu_{A_1} = \begin{cases} 0, \text{ при } x < a; \\ \frac{x-a}{b-a}, \text{при } a \le x < b; \\ 1, \text{при } b \le x \le c; \\ \frac{d-x}{d-c}, \text{при } c < x \le d; \\ 0, \text{ при } x > d. \end{cases} \tag{2.12}$$

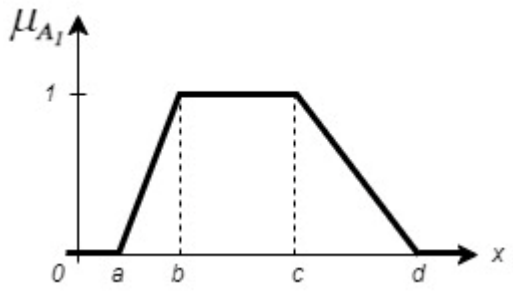


Рис. 2.3. Трапецієподібна функція належності $\mu_{A_1}(x, a, b, c, d)$

Твердження «оцінка близька до b» описується нечіткою підмножиною A_2 із трикутною функцією належності (2.13) $\mu_{A_2}(x, a, b, c)$, графік якої наведено на рисунку 2.4.

$$\mu_{A_2} = \begin{cases} 0, \text{ при } x < a; \\ \frac{x-a}{b-a}, \text{при } a \leq x < b; \\ \frac{c-x}{c-b}, \text{при } b \leq x \leq c; \\ 0, \text{ при } x > c. \end{cases} \tag{2.13}$$

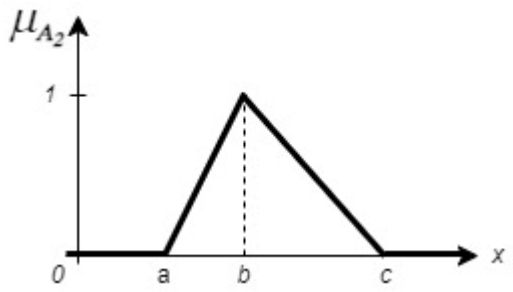


Рис. 2.4. Трикутна функція належності $\mu_{A_2}(x, a, b, c)$

Твердження «оцінка не більша за b» » описується нечіткою підмножиною A_3 з Z-образною функцією належності (2.14) $\mu_{A_3}(x, a, b)$, графік якої наведено на рисунку 2.5:

$$\mu_{A_3} = \begin{cases} 1, \text{при } x \leq a; \\ \frac{b-x}{b-a}, \text{при } a < x \leq b; \\ 0, \text{ при } x > b. \end{cases} \tag{2.14}$$

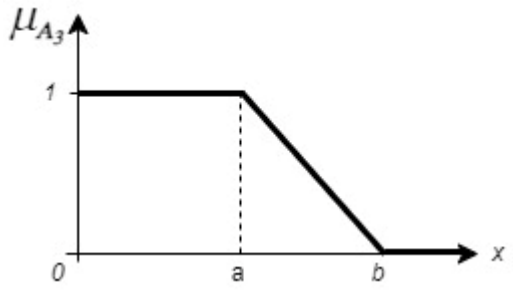


Рис. 2.5. Z-образна функція належності $\mu_{A_3}(x, a, b)$

Твердження «оцінка не менша за b» описується нечіткою підмножиною A_4 з лінійною S-образною функцією приналежності (2.15) $\mu_{A_4}(x, a, b)$, графік якої наведено на рисунку 2.6.

$$\mu_{A_4} = \begin{cases} 0, \text{при } x \leq a; \\ \frac{x-a}{b-a}, \text{при } a < x \leq b; \\ 1, \text{ при } x > b. \end{cases} \quad (2.15)$$

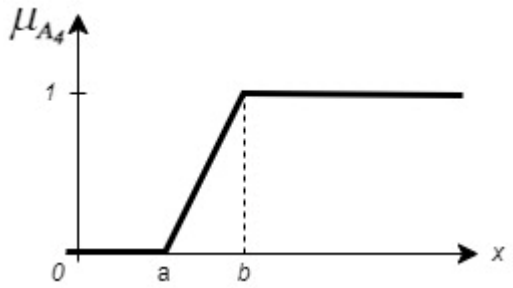


Рис. 2.6. S-образна функція належності $\mu_{A_4}(x, a, b)$

Визначивши функції належності для нечітких оцінок усіх експертів, можна визначити функцію належності колективної оцінки об'єкту, як математичне очікування (2.16):

$$\mu_A = \sum_{j=1}^{m} \mu_{A_j} \quad (2.16)$$

де μ_A – функція належності групової оцінки; μ_{A_j}– функція належності оцінки j-го експерта; m – кількість експертів.

Якщо склад групи експертів не є однорідним, експертам встановлюється ваговий коефіцієнт [86] – K_j, він враховується функцією належності колективної оцінки наступним чином (2.17):

$$\mu_A = \sum_{j=1}^{m} K_j \mu_{A_j}, \quad (2.17)$$

де μ_A – функція належності групової оцінки; μ_{A_j}– функція належності оцінки

j-го експерта; K_j – ваговий коефіцієнт j-го експерта; m – кількість експертів.

Для вирішення задачі визначення групової оцінки a може бути використано середньозважене значення (2.18):

$$a = \frac{\sum_{j=1}^{m} x_j \mu_{A_j}(x_j)}{\sum_{j=1}^{m} \mu_{A_j}(x_j)}, \tag{2.18}$$

де a – групова середньозважена оцінка; μ_A – функція належності групової оцінки; μ_{A_j}– функція належності оцінки j-го експерта; K_j – ваговий коефіцієнт j-го експерта; x_j – оцінка, що надана j-м експертом, $j = \overline{1, m}$; m – кількість експертів.

Розглянемо методи визначення оцінок експертів за допомогою лінгвістичних змінних, множини термів та нечітких чисел [73].

При людино-машиній взаємодії теорію нечітких множин також можна застосувати для опису нечітких категорій, наближених до людських понять та уявлень. Для цього використовують поняття лінгвістичної змінної, що було введено вперше разом з введенням нечіткої логіки [79, 87].

Лінгвістичні змінні дозволяють відобразити опис природньою мовою властивостей об’єктів за відсутності їх точного детермінованого опису. При цьому слід врахувати, що багато нечітких категорій, описаних лінгвістично, часто не менш інформаційні, ніж точний опис. Замість точних числових значень, експерти можуть використовувати лінгвістичні змінні для характеристики оцінок [88, 89, 90]. Наприклад, замість числа «30», у контексті кількості балів оцінки якості рівня знань студента закладу вищої освіти використовуватися лінгвістична змінна «неприйнятно» або «незадовільно».

Лінгвістична змінна – це людська суб'єктивна оцінка конкретного числового значення, що виражено природною мовою, застосовується для опису нечіткого значення, що було отримано у результаті опитування експертів.

Нечітка логіка оперує змінними, які називаються лінгвістичними. В загальному випадку лінгвістична змінна має вигляд (2.19):

$$< X, T(X), E, G, M >, \tag{2.19}$$

де X – назва лінгвістичної змінної, $T(X)$ – терм-множина лінгвістичної змінної, що визначає перелік найменувань нечітких змінних на множині X, окремі елементи цієї множини називаються термами, E – діапазон значень змінної X (числове представлення універсуму, що відповідає домену предметної галузі), G – синтаксичне правило, що описує спосіб породження термів, зокрема тих, що утворюються на основі вже визначених за рахунок використання деяких квантифікаторів як от «не», «дуже», «більш», «менш», «або», «і» та інших, що в тому числі можуть залежати від домену предметної галузі, M – семантичне правило, що визначає відповідність деякої функції належності $\mu(x)$ кожній нечіткій змінній, яка утворена за допомогою синтаксичного правила G (фактично, для кожного квантифікатора створюється окрема процедура, що породжує нову функцію належності на основі значення функції належності деякого терма) [87].

Терм-множина – це сукупність усіх можливих значень лінгвістичної змінної, а терм – це, відповідно, будь-який елемент терм-множини. Наприклад, фразу «оцінка рівня знань студента дорівнює 80 балів» можна замінити фразою «оцінка рівня знань студента –добре». У цьому контексті, лінгвістична змінна «оцінка рівня знань студента» має значення «добре», маючі на увазі, що терм «добре» грає таку ж роль, як і кількість балів 80. Терм-множина лінгвістичної змінної «оцінка рівня знань студента» може бути описана наступним чином:

«оцінка рівня знань студента» = {«відмінно», «добре», «задовільно», «незадовільно»}.

Терм-множина у загальному випадку може мати нескінченне число термів, але на практиці при наданні оцінки експертами терм-множина завжди кінцева.

Нечітке число – це нечітка змінна, що визначено на числовій осі, тобто нечітке число є нечіткою підмножиною A на множині дійсних чисел $\mathbb{R}$ с функцією приналежності [91] $\mu_{A_x}(x) \in [0; 1]$, де x – дійсне число, $x \in \mathbb{R}$.

Для лінгвістичної змінної існує поняття базової змінної – числова змінна, що відображає значення лінгвістичної змінної та може приймати значення на певній множині значень, яку називають базовою множиною чи множиною базових значень. Для лінгвістичної змінної «оцінка рівня знань студента» базовою множиною, наприклад, може бути множина {35,55,80,95}. При цьому, значення лінгвістичної змінної «добре» може бути інтерпретовано як назва нечіткого обмеження для базової змінної – це обмеження і є сенсом значення терму «добре». Для представлення числової характеристики, що є чисельним відображенням думки експерта відносно нечіткого обмеження, використовується функція належності.

Формально нечітка змінна описується трійкою (2.20):

$$< \alpha, \mathrm{X}, \mathrm{C}(\alpha) >, \tag{2.20}$$

де α – найменування змінної, $\alpha \in A$; X – універсальна множина (область визначення α); $C(\alpha)$ – нечітка множина на X, що описує обмеження на можливі значення нечіткої змінної α: $C(\alpha) = \{\mu_\alpha(\mathrm{x}) | \mathrm{x} \in \mathrm{X}\}$.

Якщо оцінка рівня знань студентів оцінюється за допомогою термів «відмінно», «добре», «задовільно», «незадовільно», при цьому мінімальна кількість балів, яку отримує студент, дорівнює 0, а максимальна 100, тоді нечітку змінну буде описано наступним чином:

$$< \text{«оцінка рівня знань студента»}, [0; 100], \mu_1(x), \mu_2(x), \mu_3(x), \mu_4(x) >.$$

Розглянемо метод побудови висновків на основі правил [92, 93, 94]. Для отримання оцінки об’єкту може бути використано систему нечіткого виведення, що представляє собою процес отримання нечітких висновків на основі нечітких умов або передумов, що представляють собою інформацію про властивості об'єкта. У загальному випадку процес нечіткого виведення складається з наступних етапів:

1) формування бази правил;
2) фазифікація вхідних даних;
3) агрегування передумов;
4) активація (композиція) підзаключень;
5) акумулювання висновків;
6) дефазифікація вихідних змінних.

База правил містить формальне уявлення емпіричних знань експертів про властивості об'єкту. Формальне представлення надається у вигляді нечітких продукційних правил (правил-продукції), що представляють знання експертів у вигляді лінгвістичного правила «ЯКЩО-ТО» [94]. Це правило представляє собою конструкцію типу (2.21):

$$R = \text{«ЯКЩО } x \in A\text{, ТО } y \in B\text{»}, \tag{2.21}$$

де «$x \in A$» – нечітке висловлювання, передумова правила (посилка чи антецедент); «$y \in B$» – нечітке висловлювання, висновкок правила (дія чи консеквент).

Виглядає база правил як структурований текст, наприклад наступним чином:

Правило 1. ЯКЩО «умова A_1» або «умова A_2» ТО «висновок B_1»

Правило 2. ЯКЩО «умова A_1» та «умова A_3» ТО «висновок B_2»

і т.д.

Залежно від кількості передумов та висновків база правил може бути представлена структурою одного з наступних типів [95, 96, 97]:

– SISO – структура з одним входом та одним виходом;

– MISO – структура, з багатьма входами та одним виходом;

– MIMO – структура з багатьма входами та багатьма виходами .

База даних системи систему нечіткого виведення містить визначені функції приналежності, що використовуються в продукційних правилах. Фазифікація, або введення нечіткості, передбачає встановлення відповідності між вхідним чисельним значенням та значенням функції належності, що відповідає його терму. В процесі агрегування отримані після фазицікації значення функцій належності термів використовуються для визначення агрегованої ступені істинності за кожної передумови кожного правила системи. В процесі активації визначається функція належності до підзаключень, використовуючи методи нечіткої композиції. В процесі акумулювання виконується пошук функції належності для кожної з вихідних лінгвістичних змінних, заданих сукупністю правил. Наприкінці, дефазифікація виконує перехід від функції належності вихідної лінгвістичної змінної до її чітких значень.

Компоненти системи нечіткого виведення можуть бути реалізовані по різному та визначати алгоритм нечіткого виведення. Серед найбільш поширених на теперішній час алгоритмів можна вказати алгоритми Мамдані, Ларсена, Цукамото, Такагі–Сугено [92, 98, 99].

2.3 Підходи до обробки текстових даних природною мовою

Основою будь-якого автоматизованого процесу, що забезпечує діалог з користувачем його природньою людською мовою, є необхідність правильного розуміння мови користувача та сенсу його запиту. Одною з головних проблем сучасної рекомендаційної системи є проблеми представлення та отримання значення лінгвістичної інформації, що надано користувачем. Питанням

вилучення інформації шляхом представлення та інтерпретації письмової та усної людської мови займається розділ штучного інтелекту у сфері машинного навчання, що отримав назву «обробка природньої мови» (Natural Language Processing, або NLP) [100, 101, 102]. На теперішній час методи та інструменти NLP використовують такі додатки як: голосові помічники, чат-боти, автоматизовані перекладачі, аналізатори настроїв та, звичайно, пошукові системи [103, 104].

2.3.1 Методи попередньої обробки даних та підходи до векторизації даних

Першим етапом обробки тексту природньою мовою є збір текстових даних, які можуть бути розміщені на сторінках веб-сайтів, у документах, базах даних та інших джерелах. Після збору текстових даних проводять їх попередню обробку, з метою очищення даних від непотрібних елементів: символів, незначущих слів та приведення їх до нормальних або усічених форм. Попередня обробка запиту та тексту надає їм одноманітності, забезпечує високу продуктивність і достовірність пошуку подібності та відповідності, що підвищує високу якість функціонування рекомендаційних систем [105, 106, 107].

Розглянемо основні методи попередньої обробки текстових даних, які можна використати при вирішенні задач рекомендаційних систем.

Токенізація – процес поділу тексту на компоненти. Розрізняють токенізацію по реченнях, коли текст поділяється на такі компоненти, як речення, та токенізацію по словах, коли текст поділяється на такі компоненти, як прості чи складені слова або словосполучення. Обидві задачі виглядають простими, але є нетривіальними. У багатьох мовах ознака закінчення речення – крапка – також може використовуватися у скороченнях (назв, або імен) чи як частина багато крапки, а символ пробілу хоч і є основним роздільником слів, але, наприклад, є ситуації, коли слова пишуться через дефіс.

Вилучення стоп-слів – виключення з вхідного корпусу текстів слів, що не несуть смислового навантаження, але при цьому, як правило, мають доволі високу частоту зустрічання в тексті. Найчастіше до стоп-слів відносяться прийменники, цифри, займенники, тощо, однак універсального словника стоп-слів не існує і такий словник може бути сформовано відповідно до поставленої задачі. В результаті вилучення стоп-слів зменшується об’єм тексту для опрацювання та знижається його зашумленість [108, 109].

Після вилучення зайвого шуму проводять нормалізацію тексту – приведення слів тексту до початкових форм або до основи: іменники мають початкову форму у однині називного відмінку, прикметники у однині чоловічого роду називного відмінку і т.д. [110]. Основними підходами нормалізації тексту є стемінг та лематизація. Стемінг передбачає приведення слова до його основи і використовує грубий евристичний підхід, при приведенні від слова відсікаються суфікси та закінчення [111, 112]. Отриманий стемером результат не може вважатися морфологічною основою [113]. Лематизація – процес приведення слова до смислової канонічної форми, що називається лемою – це саме та форма, що заноситься до словнику мови [114]. Для підвищення рівня точності лематизації використовують визначення для слова частини мови, оскільки для різних частин речі можуть використовуватися різні правила слово формування. У порівнянні зі стемінгом лематизація має вищу точність, але уступає йому у швидкості обробки. Для вилучення семантичної інформації іноді виконують морфологічну сегментацію – поділ слів на морфеми (основні одиниці слова) [115, 116]. Морфемою називають найменшу одиницю слова, яка несе значення. Прості слова складаються з однієї морфеми, а складні можуть мати декілька, наприклад, слово «енергозбереження» складається з двох морфем «енергія» та «збереження».

Попередня обробка тексту, в загальному випадку, не потребує реалізацій, оскільки наразі у сфері обробки природньої мови існує велика

кількість інструментарію для різних мов програмування, що функціонально забезпечені усіма методами для попередньої обробки текстів [117, 118, 119].

Для опрацювання даних алгоритмами машинного навчання після процесу нормалізації тексту виконують його векторизацію – переведення тексту у числовий вигляд, з яким можуть працювати алгоритми. Існують декілька підходів до векторизації текстів, розглянемо найбільш поширені з них.

Мішок слів (Bag of Words) дозволяє представити текст у вигляді числового вектору входження кожного слова [120, 121, 122]. При цьому способі розглядається лише множина слів (токенів), а порядок та структура ігноруються. За логікою цього способу два документа вважаються однаковими, якщо складаються з однакових токенів.

У термінах NLP кожне окреме речення можна виступати документом, а сукупність документів – корпусом. Наприклад, маємо у корпусі C три документи D_1, D_2, D_3, що наведено у таблиці 2.4:

Таблиця 2.4 – Текстовий корпус

Документи	Вміст документу
D_1	*Завдання на лабораторну роботу*
D_2	*Методичні вказівки до виконання лабораторних робіт*
D_3	*Лекції, методичні вказівки до практичних робіт, методичні вказівки до лабораторних робіт та інші методичні матеріали*

Для побудови мішка слів на першому кроці створюється словник W унікальних токенів у корпусі (тексти корпусу повинні пройти певну попередню обробку, в тому числі, етапи виділення токенів та очищення від стоп-слів):

W = { *«завдання», «лабораторний», «робота», «методичний», «вказівка», «виконання», «лекція», «матеріал» «практичний»*}.

На другому кроці будується таблиця входження слів зі словника для кожного документа. При чому для оцінки входження токенів у документ використовується декілька підходів:

1) бінарний – використовується лише інформацію про входження, або не входження токену у документ (табл. 2.5, заповнюється лише значеннями 0 та 1);

2) кількісний – використовується інформація про кількість входжень токенів у документ (табл. 2.6, заповнюється відповідними знаннями кількості для кожного токену);

3) частотний – використовується інформація як часто кожний токен зустрічається у тексті, по відношенню до загальної кількості слів у тексті (табл. 2.7, заповнюється значеннями частот для кожного токену);

Таблиця 2.5 – Таблиця бінарного входження слів за документами

Документи	*«завдання»*	*«лабораторний»*	*«робота»*	*«методичний»*	*«вказівка»*	*«виконання»*	*«лекція»*	*«матеріал»*	*«практичний»*	*«інший»*
D_1	1	1	1	0	0	0	0	0	0	0
D_2	0	1	1	1	1	1	0	0	0	0
D_3	0	0	1	1	1	0	1	1	1	1

Таблиця 2.6 – Таблиця кількісного входження слів за документами

Документи	*«завдання»*	*«лабораторний»*	*«робота»*	*«методичний»*	*«вказівка»*	*«виконання»*	*«лекція»*	*«матеріал»*	*«практичний»*	*«інший»*
D_1	1	1	1	0	0	0	0	0	0	0
D_2	0	1	1	1	1	1	0	0	0	0
D_3	0	1	2	3	2	0	1	1	1	1

Таблиця 2.7 – Таблиця частотного входження слів за документами

Документи	*«завдання»*	*«лабораторний»*	*«робота»*	*«методичний»*	*«вказівка»*	*«виконання»*	*«лекція»*	*«матеріал»*	*«практичний»*	*«інший»*
D_1	$1/3$	$1/3$	$1/3$	0	0	0	0	0	0	0
D_2	0	$1/5$	$1/5$	$1/5$	$1/5$	$1/5$	0	0	0	0
D_3	0	$1/12$	$1/6$	$1/4$	$1/6$	0	$1/12$	$1/12$	$1/12$	$1/12$

До недоліків методу можна віднести отримання найвищої оцінки токенами, що зустрічаються у документі частіше, однак на практиці не обов’язково частота цих слів буде значущою для визначення характеристик тексту, наприклад, тематична спрямованість тексту чи перелік ключових слів предметної галузі.

TF-IDF(Term Frequency – Inverse Document Frequency) – спосіб використовує оцінку ваги слова документа з точки зору документа, вирішуючи проблему необґрунтованої ваги токенів при використанні мішка слів. За логікою цього способу якщо токен і зустрічається часто у одному документі, але зустрічається зрідка, або взагалі не зустрічається у інших документах, то його вага невелика [123, 124, 125].

TF – відносна частота токену у корпусі, для кожного токену документа обчислюється як відношення кількості входжень токену у документ до загальної кількості токенів у словнику (2.22):

$$TF_{ij} = \frac{d_{ij}}{|W|}, \tag{2.22}$$

де TF_{ij} – відносна частота токену у корпусі; d_{ij} – кількість входжень токену i у документ j, $i = \overline{1, |W|}$, $j = \overline{1, |C|}$; $|W|$ – потужність множини токенів (кількість токенів у корпусі); D_k – документ, $k = \overline{1, |C|}$;

IDF (Inverse Document Frequency) – зворотня відносна частота токену у корпусі, для кожного токену документа обчислюється як відношення загальної кількості документів до кількості документів де зустрічається токен (2.23):

$$IDF_{ij} = log\ (\frac{|C|}{|\{D_k|d_{ij} \in D_k,\}|}), \quad (2.23)$$

де IDF_{ij} – зворотна відносна частота токену у корпусі; $|C|$ – потужність множини корпусу (кількість документів); D_k – документ, $k = \overline{1,|C|}$; d_{ij} – кількість входжень токену i у документ j, $i = \overline{1,|W|}$, $j = \overline{1,|C|}$.

Остаточний результат, оцінка ваги токену у документі визначається множенням частоти TF на частоту IDF (2.24):

$$TFIDF_{ij} = TF_{ij} \cdot IDF_{ij}, \quad (2.24)$$

де TF_{ij} – відносна частота токену у корпусі, $i = \overline{1,|W|}$, $j = \overline{1,|C|}$; IDF_{ij} – зворотна відносна частота токену у корпусі, $i = \overline{1,|W|}$, $j = \overline{1,|C|}$.

В результаті отримуємо таблицю 2.8 оцінок токенів корпусу за документами.

Таблиця 2.8 – Оцінки токенів корпусу за документами за метрикою TF-IDF

Документи	*«завдання»*	*«лабораторний»*	*«робота»*	*«методичний»*	*«вказівка»*	*«виконання»*	*«лекція»*	*«матеріал»*	*«практичний»*	*«інший»*
D_1	0,142	0,3	0,3	0	0	0	0	0	0	0
D_2	0	0,3	0,3	0,242	0,242	0,142	0	0	0	0
D_3	0	0,3	0,6	0,725	0,483	0	0,142	0,142	0,142	0,142

Недоліком методу є великий обсяг даних для опрацювання, коли вектор слів може містити тисячі токенів. У випадку коли кожен документ містить лише малу кількість слів зі словника, вектори документів та матриця корпусу будуть розрядженими, тобто містити переважно нулі. Розряджені вектори та матриці вимагають великої кількості пам'яті та обчислювальних ресурсів.

Мішок слів та TF-IDF не зберігають інформацію ні про структуру тексту документу ні про його контекст. Ця проблема може бути вирішена вирішити за допомогою комбінації послідовності токенів, або N-грами, де позначення N описує кількість токенів у грамі. Так, уніграма буде складатися з одного токену, біграма з двох і так далі. Спосіб векторизації N-грамами спрощує визначення контексту документа за рахунок збереження певних послідовностей токенів довжиною N. Відповідно, чим довші будуть визначені грами, тім більше контекстної інформації вони зберігають. Наприклад, для документу «Лекції, методичні вказівки до практичних робіт, методичні вказівки до лабораторних робіт та інші методичні матеріали» доречніше було б оперувати не токенами, а такими грамами як: «методичні вказівки», «вказівки до практичних робіт», «вказівки до лабораторних робіт», «методичні матеріали» тощо. Для вилучення N-грам з документу існує багато алгоритмів, серед поширеніших можна вказати Алгоритм Лемпеля-Зіва-Велча (LZW), алгоритми суфіксного дерева та суфіксного масиву, алгоритм інвертованого індексу [126, 127, 128].

Вектори слів можна легко кластеризувати за допомогою алгоритмів кластеризації [129, 130]. Отримані кластери можна використовувати як функції в алгоритмах навчання, що працюють з дискретними функціями.

Розглянуті векторні представлення документів дозволяють виконувати порівняння документів шляхом визначення відстані між ними. Якщо два документи представлені в n-вимірному просторі векторами p та q, то відстано між ними може бути обчислена з використанням однієї з нижченаведених метрик [131, 132]:

– Евклідова відстань (Euclidean Distance) – вимірює пряму відстань між двома векторами у *n*-мірному просторі (2.25):

$$d_e(p,q) = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}; \quad (2.25)$$

– Манхеттенська відстань (Manhattan Distance) – відстань між двома векторами у *n*-мірному просторі, що вимірюється по осях, розташованих під прямим кутом (2.26):

$$d_m(p,q) = \sum_{k=1}^{n}|p_k - q_k|; \quad (2.26)$$

– відстань Чебишева (Chebyshev Distance) – відстань між двома векторами у *n*-мірному просторі, що представляє найбільшу різницю векторних координат (2.27):

$$d_{ch}(p,q) = max_{k=1..n}(|p_k - q_k|); \quad (2.27)$$

– косинусна схожість (Cosine Similarity) – коефіцієнт схожості між двома не нульовими векторами у предгільбертовому просторі, що обчислюється як косинус кута між ними (2.28):

$$\cos_sim(p,q) = \frac{\sum_{k=1}^{n} p_k q_k}{\sqrt{\sum_{k=1}^{n} p_k^2}\sqrt{\sum_{k=1}^{n} q_k^2}}. \quad (2.28)$$

Як правило, частіше за все використовують косинусну відстань, оскільки результати близькості векторів текстів схожі з експертними оцінками семантичної схожості текстів [133, 134].

2.3.2 Методи Natural Language Processing для визначення ключових елементів тексту та семантичних зв’язків

Для визначення ключових елементів тексту та семантичних зв’язків в тексті найчастіше застосовуються методи розпізнавання іменованих об’єктів.

Розпізнавання іменованих об’єктів (Named Entity Recognition, NER) – задача в області обробки природньої мови, мета якої вилучення інформації про іменовані об’єкти за попередньо визначеними категоріями [135, 136, 137]. До категорій іменованих об’єктів відносять імена осіб, організацій, місця розташування, грошові значення тощо. Для обробки природньої мови розуміння цих сутностей має велике значення, оскільки вони часто є важливою інформацією у тексті.

Процес розпізнавання іменованих об’єктів можна представити двома основними етапами.

1. Ідентифікація суб'єкта. За допомогою лінгвістичних правил або статистичних методів виявляються потенційні іменовані сутності. Процес базується на розпізнаванні шаблонів, характерних для таких сутностей, наприклад, великі літери в іменах та прізвищах, або певних форматах, наприклад, формат запису дати.

2. Класифікація суб'єкта. Після ідентифікації об’єкти класифікують у попередньо визначені класи, наприклад, «організація» або «особа». Для цього використовуються моделі машинного навчання, навчених на позначених наборах даних.

Існують багато методів для вирішення задач розпізнавання іменованих об’єктів, розглянемо основні найбільш популярні підходи до їх реалізації.

Методи на основі побудованих вручну правил. Ідентифікація виконується на основі лінгвістичних моделей, регулярних виразів або словників. Методи можуть проблемними, обробляючі великі та різноманітні набора даних через не гучність попередньо визначених правил, хоча у певних сферах їх використання більш ніж достатньо.

Статистичні методи використовують такі моделі, як приховані марковські моделі (Hidden Markov models) [138] або умовні випадкові поля (Conditional random fields) [139]. Методи визначають іменовані сутності на основі ймовірностей, отриманих із даних навчання. Методи використовують для завдань із достатньою кількістю позначених наборів та результат їх роботи сильно залежить від якості даних на яких їх було навчено.

Методи машинного навчання використовують для вирішення задачі такі алгоритми, як дерева рішень або опорні векторні машини [140]. Ці методи добре працюють з великими наборами даних та складними шаблонами, тому знайшли застосування у багатьох системах розпізнавання іменованих об’єктів. Однак, такі методи потребують великого об’єму позначених даних і можуть бути вимогливими до обчислень.

Методи глибокого навчання використовують нейронні мережі. Найбільш популярними наразі є рекурентні нейронні мережі (Recurrent neural network) [141] та трансформери (Transformers) [142, 143] завдяки їх здатності моделювати довготривалі залежності в тексті. Методи гарно працюють із навчальними даними великого об’єму, але потребують значної обчислювальної потужності.

Гібридні методи поєднують підходи на основі правил, статистики та машинного навчання та є найкращими при розпізнаванні іменованих об’єктів із різноманітних джерел, завдяки гнучкості використання кількох методів.

2.3.4 Моделі та методи обробки текстових даних на основі нейронних мереж

Нейронні мережі забезпечують потужний механізм навчання, який є дуже привабливим для вирішення проблем обробки природної мови. Основним компонентом нейронних мереж є використання шару вкладення, відображення дискретних символів у безперервні вектори у відносно низьковимірному просторі. Вкладеннями, або embedding, називають векторне представлення слів, яке було отримано у результаті навчання [144]. Під час

вкладання слів вони перетворюються з окремих слів на математичні об’єкти, якими можна оперувати, зокрема, можна прирівняти відстань між векторами вкладень. Векторне представлення слів використовується мережею, як частина процесу навчання. Мережу також можна навчити комбінувати вектори слів у спосіб, який є корисним для прогнозування. Ця можливість певною мірою усуває проблеми дискретності та розрідженості даних.

Рекурентна нейромережева мовна модель (RNNLM) вивчає одночасно представлення кожного слова та функцію ймовірності для сусідніх послідовностей слів, приймає на вхід векторні представлення *n* попередніх слів і, умовно, може розуміти семантику речення. Навчання моделі базується на алгоритмі безперервного мішка слів [145, 146]. Сусідні від слова, що розглядається, вважаються контекстами, передаються на вхід нейронної мережі, яка передбачає центральне слово. Отримані вектори стискаються, об’єднуються та передаються в прихований шар нейронної мережі, де спрацьовує *softmax* функція активації, яка й визначає, які сигнали пройдуть далі по мережі. Початкова версія моделі ґрунтувалася на нейромережах прямого поширення, де сигнал йде безпосередньо від вхідного шару мережі до вихідного. Пізніше була запропонована модель у вигляді рекурентних нейронних мереж (RNN). Компания Google надає доступ до готової навченої моделі для багатьох мов, яка використовує три приховані шари нейронної мережі прямого поширення та 128 векторних вкладень. [147].

До переваг моделі можна віднести доволі швидке навчання, прості вкладення, яких достатньо для використання у простих додатках та доступність попередньо навченої моделі для багатьох мов. Недоліками моделі можна вважати не врахування моделлю довгострокових залежностей.

Модель WordToVec – загальна назва для сукупностей моделей на базі штучних нейронних мереж, які призначені для отримання векторних вкладень слів природною мовою. Концепції методу, окрім отримання вкладень, успішно використовуються при розробці рекомендаційних систем [148, 149]. В моделі рекомендацій цю технологію використовують такі відомі компанії як Airbnb,

Alibaba або Spotify [150, 151]. Підхід заснований гіпотезі локальності: «слова, що зустрічаються в однакових оточеннях, мають близькі значення», де близькість розуміється доволі широко – поруч можуть стояти лише слова, що поєднуються. Наприклад, словосполучення «методичні вказівки» може існувати, слова поєднуються, але «методичні яблука» не може, бо слова не поєднуються. Word2vec виконує прогнозування на основі контекстної близькості цих слів, навчаючи просту нейронну мережу. Тому, щоб отримувати ефективні результати його роботи та підвищити якість передбачень, навчання повинно проводитися на великих корпусах.

За узагальненим принципом, модель Word2vec приймає корпус тексту, генерує словник корпусу, та обчислює векторне представлення слів, навчаючись на вхідних текстах. Алгоритм передивляється кожен токен у тексті, який є центральним токеном t_{ce} і контекстним токеном t_{co}. Далі використовується схожість векторів токенів для t_{ce} і t_{co}, щоб розрахувати ймовірність t_{co} при заданому t_{ce}, або навпаки. Регулювання вкладень продовжується для максимізації цієї ймовірності. Отримані в результаті векторні представлення слів дозволяють обчислювати «семантичну відстань» між словами та можуть бути використані для обробки природної мови та машинного навчання.

WordToVec представляється двома модифікаціями та алгоритмами для навчання: CBOW (Continuous Bag of Words) та Skip-Gram [152, 153].

Основна ідея Skip-Gram – передбачати навколишні слова, використовуючи поточне слово. Алгоритм розглядає контекстне вікно, що містить k послідовні слова, пропускає одне слово і навчає нейронну мережу, що містить усі слова, крім пропущеного – саме його алгоритм і намагається передбачити. Таким чином, якщо 2 слова періодично мають схожий контекст у корпусі, ці слова матимуть близькі значення векторів.

Ідея Continuous Bag of Words передбачити поточне слово, виходячи з навколишнього контексту. Алгоритм розглядає багато документів у корпусі. Кожно разу, коли алгоритм бачить слово, він береться сусіднє до розглянутого

слова, подаються на вхід нейромережі, яка передбачє слово у центрі цього контексту. У випадку тисяч таких контекстних слів та центрального слова, отримується новий екземпляр для нейромережі. Нейромережа тренується і на виході закодованого прихованого шару представляє вкладення для певного слова. Те саме відбувається, якщо нейромережа тренується на великій кількості речень і слів у схожому контексті приписуються схожі вектори.

Існує та є у вільному доступі попередньо навчена модель, яка може бути імпортована за допомогою бібліотеки Gensim [154].

До переваг моделі можна віднести просту архітектуру, швидке навчання та генерацію вкладень, можливість використання моделі у інших областях. Недоліками моделі можна вказати відсутність інформації про контекст, у якому використовується слово, бо навчається система лише на рівні слів, за тою ж причиною модель не враховує, що слово може мати різний значення, відповідно до контексту використання.

Модель GloVe намагається вирішити проблему ефективного використання статистики збігів, використовуючи стохастичний градієнтний спуск для мінімізації різниці між добутком логарифмом ймовірності спільної появи слів та векторів слів. Набуті моделлю вкладення представляють собою підструктури векторного простору слів, що зв'язують разом різні, але пов’язані між собою загальним контекстом слова, наприклад поштовий код міста та його назву [155]. У модель Word2Vec та, наскільки часто зустрічаються спільно слова не має впливу, а лише допомагає формувати додаткові навчальні вибірки, В порівнянні з нею, модель GloVe, групує вектори слів разом на основі їхньої глобальної схожості, тобто враховує спільну зустрічальність слів, не спираючись лише на контекстну статистику. Навчена модель GloVe знаходиться у відкритому доступі.

До переваг моделі можна віднести швидкість, формування осмислених вкладень, модель краща за Word2Vec, додаючи частоту слів та надає кращі результати на контрольних завданнях. Недоліками моделі можна вказати все

ще недостатній рівень даних об контексті та погану роботу з рідкісними та невідомими словами.

FastText – це бібліотека з відкритим похідним кодом для представлення та класифікації тексту, була розроблена командою дослідження штучного інтелекту Facebook (FAIR). Модель призначена для ефективної обробки великих обсягів текстових даних і надає інструменти для класифікації тексту, представлення слів і обчислення подібності тексту [156]. У своїй основі FastText використовує концепцію вбудовування слів, щільними векторними представленнями у неперервному векторному просторі. Вбудовування слів надає можливість фіксувати семантичні та синтаксичні зв'язки між словами на основі їх властивостей розподілу у певному текстовому корпусі. Для векторизації слів модель використовуються одночасно Skip-Gram, негативне сем плювання та алгоритм безперервного мішка. До основної моделі Word2Vec додано модель символьних N-грам, коли кожне слово представляється композицією декількох послідовностей символів певної довжини N, тож результати класифікації підходять краще до слів з невисокою частотою зустрічальності. На відміну від моделей Word2Vec та Glove, FastText добре генерує вкладення до рідких та невідомих слів. Готова модель FastText також знаходиться у відкритому доступі.

До переваг моделі можна віднести її просту архітектуру: один вхід, один прихований шар та один вихід, добре працює на побудові вкладень для рідкісних та невідомих слів Недоліками моделі можна вказати навчання на рівні слів, тож не містить інформації про речення або контекст, модель не враховує сумісну зустрічалість.

Трансформери (Transformers) – відносно нова архітектура на базі глибоких нейроних мереж, яка з’явилася у 2017 році. Трансформери націлені на задачу розв'язання текстових послідовностей та обробку далекодіючих залежностей. На сьогодні час технологія трансформерів вважається найбільш просунутою у галузі обробки природньої мови. Моделі-трансформери

використовуються при перекладанні текстів, генеруванні текстів, та навіть у генеруванні програмного коду.

Архітектура трансформерів містить два головних компоненти: кодувальник (енкодер), що конвертує вхідний текст у векторне представлення, та декодувальник (декодер, який формує текст у вигляді нової послідовності.

Рекурентні нейронні мережі, опрацьовуючи великі документи просуваються послідовно, втрачаючи зміст. Прикладом може бути ситуація, коли при використання рекурентних мереж на прикінці обробки документу модель втрачає інформацію властивості об’єкту (наприклад, про стать людини). Також, послідовна обробка, яку використовують рекурентні мережі не дозволяє робити процес навчання паралельним, що призводить до неможливості його прискорення. На відміну від рекурентних нейронних мереж, трансформери не обробляють послідовності по порядку. Наприклад, якщо вихідні дані — текст, їм не потрібно обробляти кінець пропозиції після обробки початку. Завдяки цьому таку нейромережу можна розпаралелити та навчити значно швидше.

Окрім того, трансформери використовують такі іновіційні концепції, як: позиційні енкодери (Positional Encodings), увага (Attention) та самоувага (Self-Attention) [157].

Серед найбільш популярних моделей трансформерів можна вказати наступні:

1) BERT (Bidirectional Encoder Representations from Transformers) – мовна модель, що представляє собою кодувальник, який використовує механізм «уваги» для виявлення контекстних відносини між словами у тексті, технічним нововведенням у BERT є застосуванням двонаправленого навчання [158, 159];

2) GPT (Generative Pre-trained Transformer) – велика мовна модель, з багатьма параметрами, здатна запам'ятовувати і аналізувати інформацію, створюючи на її основі зв'язний та логічний текст, використовуючи зв’язок

між окремими словами, здатна обчислювати найбільш релевантну послідовність речень та слів в них [160];

3) XLNet (eXtreme Language Model) – авторегресійна мовна модель, що використовує рекурентну архітектуру та функцію Cross-Layer Attention для обробки контекстних залежностей, виконує обчислення ймовірності спільної зустрічі послідовності слів та ймовірності для кожного слова лише за умови наявності всіх інших слів у реченні [161].

Перевагами архітектури трансформерів можна визначити підтримку врахування дострокових контекстів для слів, можливість проводити паралельне швидке навчання моделі, можливість використання у багатьох сферах обробки природньої мови, та здатність обробляти великі корпуси тексту. До недоліків можна віднести неефективність при обробці малих обсягів даних, яких недостатньо для якісного навчання, неможливість використання на послідовних даних та потреби у великих обчислювальних ресурсах

Висновки до розділу 2

В другому розділі проведено аналіз моделей та методів обробки даних рекомендаційної системи.

Оскільки параметри об’єктів інтересу, користувачів та їх взаємодії в рекомендаційній системи в сфері освіти можуть бути як кількісними, так і якісними показниками, було проведено аналіз існуючих способів представлення даних з використанням шкал та їх застосування при подальшій оцінці об’єктів різними способами. Досліджено способи розрахунків колективних оцінок, які можуть використовуватись, як результати розрахунків по матриці колаборативної фільтрації. Досліджено питання узгодженості наданих експертних оцінок.

Зважаючи на те, що оцінні судження користувачам доволі часто зручніше виконувати з використанням лінгвістичних понять, аніж числових, було запропоновано використовувати нечітку логіку для представлення таких

оцінок. Розглянуто поняття нечіткої множини, проведено аналіз методів для побудови функцій належності та метод визначення оцінок експертів за допомогою функцій належності. Визначено переваги використання нечіткої логіки при представленні даних рекомендаційних систем та наведено приклади опису понять предметної галузі сфери освіти з використанням нечітких змінних. Для забезпечення підтримки прийняття рішень в нечіткій логікі використовується метод побудови висновків на основі правил.

Елементи даних, що описують об’єкти інтересу в сфері освіти в більшості випадків включають текстову інформацію різного обсягу і потребують аналізу в їх лінгвістичній формі. Це обумовило розгляд в розділі 2 підходів до обробки текстових даних природною мовою. Огляд методів попередньої обробки дозволив визначити доцільність застосування кожного з них в процесі підготовки текстів до основного етапу обробки. Також досліджено типові моделі та методи, що забезпечують обробку статистичної та семантичної складової текстової інформації природної мови. Розглянуті в розділі 2 моделі та методи лягли в основу моделі даних рекомендаційної системи на основі нечіткої логіки та методу формування рекомендацій по об’єктам інтересу з використанням гібридного підходу, які розглянуто в розділі 3.

РОЗДІЛ 3.
РОЗРОБКА МЕТОДИКИ ПОБУДОВИ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ ДЛЯ СФЕРИ ОСВІТИ НА ОСНОВІ ГІБРИДНОГО ПІДХОДУ

3.1 Визначення характеристик рекомендаційної системи для сфери освіти

3.1.1 Визначення задач користувачів рекомендаційної системи для сфери освіти

Рекомендаційна система в сфері формальної освіти повинна бути частиною загальної системи управління освітнім процесом в межах закладу освіти та може задовольняти потреби здобувачів освіти при підтримці прийняття рішень у наступних задачах:

– побудова індивідуальної освітньої траєкторії за рахунок обирання певних дисциплін блоку вільного вибору;

– підбір навчальної літератури (посібників, підручників, методичних матеріалів тощо) як для вивчення дисциплін навчального плану, так і для самоосвіти;

– формування списків курсів інформальної освіти з урахуванням інтересів, спеціалізації чи інших параметрів здобувача освіти чи навчальної дисципліни, яку він вивчає;

– вибір тематики кваліфікаційної роботи, наукової роботи, наукового, практичного чи іншого студентського проєкту та керівника роботи.

В сфері інформальної освіти (самоосвіти) рекомендаційна система повинна бути інтегрована у відповідну платформу, на якій розміщуються курси, та забезпечувати підтримку вибору навчального курсу відповідно до інтересів слухача або до рейтингу навчального курсу [162, 163].

В рамках формальної освіти студент може потребувати наступних видів рекомендацій:

– рекомендація дисципліни з каталогу вибіркових дисциплін при формуванні індивідуальної освітньої траєкторії студента;

– рекомендація щодо можливості/доцільності використання того чи іншого джерела з бібліотеки закладу освіти при вивченні дисципліни чи для самоосвіти;

– рекомендація щодо вибору теми (або тематики дослідження) кваліфікаційної роботи чи наукової роботи, проєкту тощо;

– рекомендація при виборі потенційного керівника кваліфікаційної роботи, наукової роботи, проєкту тощо;

– рекомендація щодо використання зовнішніх курсів для організації самоосвіти (за умови, що відповідна інформація про ці курси присутня в базі рекомендаційної системи або внесена до баз даних закладу освіти).

Слід зауважити, що остання рекомендація щодо використання зовнішніх курсів передбачає значні витрати людських ресурсів та часу на формування каталогу подібних курсів з боку закладу освіти. При цьому курси зазвичай розташовуються не на одній платформі, а їх релевантність з точки зору корисності для освітнього процесу зазвичай визначається безпосередньо викладачем – фактично, його експертна думка замінює в цьому випадку роботу рекомендаційної системи. Крім того, перешкодами для формування деякого каталогу курсів виступає значна мінливість курсів (поява нових курсів, закриття існуючих курсів, зміна менторів/лекторів тощо). Це значною мірою ускладнює підтримку каталогу зовнішніх курсів в актуальному стані.

Узагальнена модель об’єктів інтересу користувачів в сфері освіти для користувачів групи «Здобувачі освіти» (студенти) може бути представлена схемою на рис.3.1.

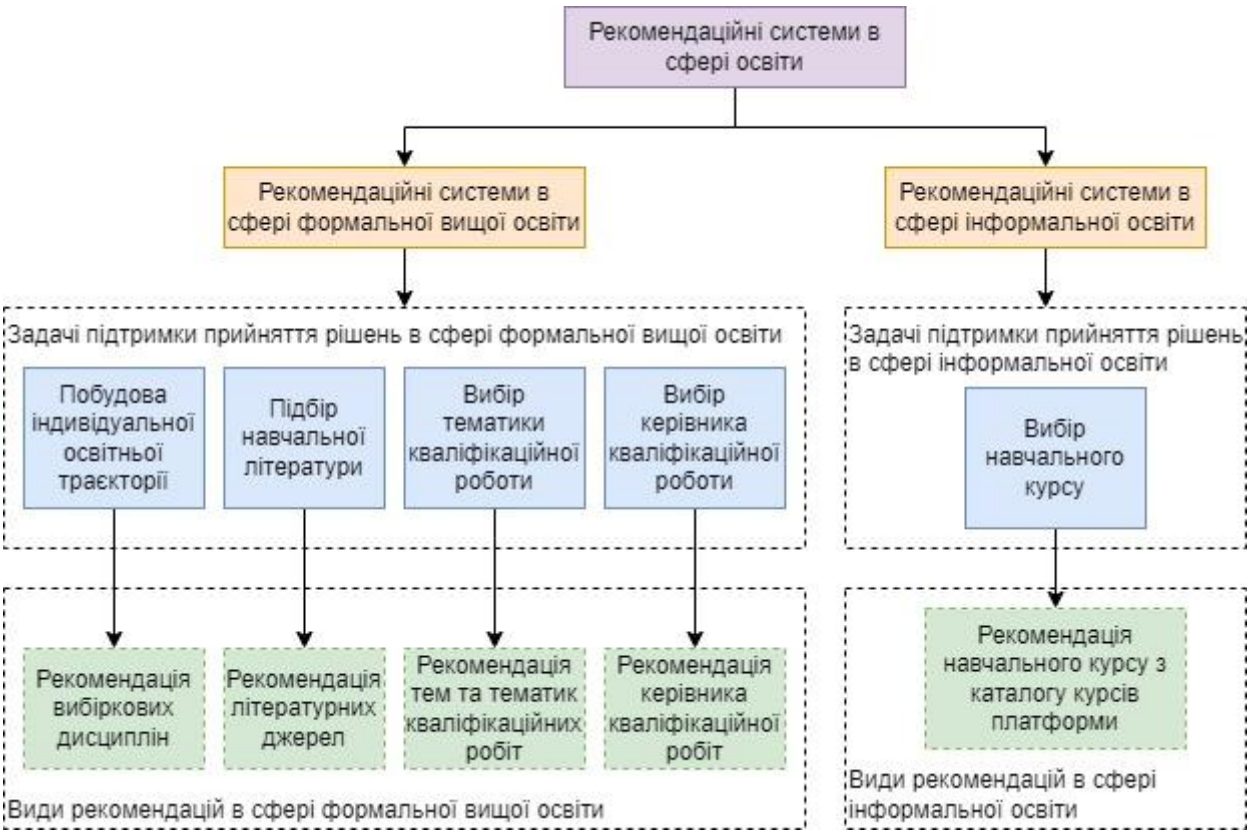


Рис.3.1. Узагальнена модель об’єктів інтересу користувачів в сфері освіти для користувачів групи «Здобувачі освіти»

3.1.2 Визначення характеристик об'єктів інтересу

Об’єктами інтересу здобувача освіти в рекомендаційній системі в рамках формальної освіти виступають наступні:

– дисципліни з каталогу вибіркових дисциплін;

– джерела з бібліотеки закладу освіти;

– теми (або тематики досліджень) кваліфікаційних робіт;

– потенційні керівники кваліфікаційних, наукових та інших робіт;

– зовнішні курси для організації самоосвіти (при наявності інформації про ці курси в базах даних рекомендаційної системи).

Перш, ніж визначати характеристики об’єктів інтересу здобувача, проведемо аналіз джерел, на основі яких може бути сформовано інформацію про об’єкти інтересу.

Ресурси для визначення параметрів об'єктів інтересу здобувача в рамках формальної освіти включають в себе наведені на рис.3.2 і включають наступні:

– силабуси дисциплін;
– бібліотека закладу освіти;
– каталог вибіркових дисциплін;
– база даних кваліфікаційних робіт;
– база даних керівників кваліфікаційних робіт;
– результати опитувань.

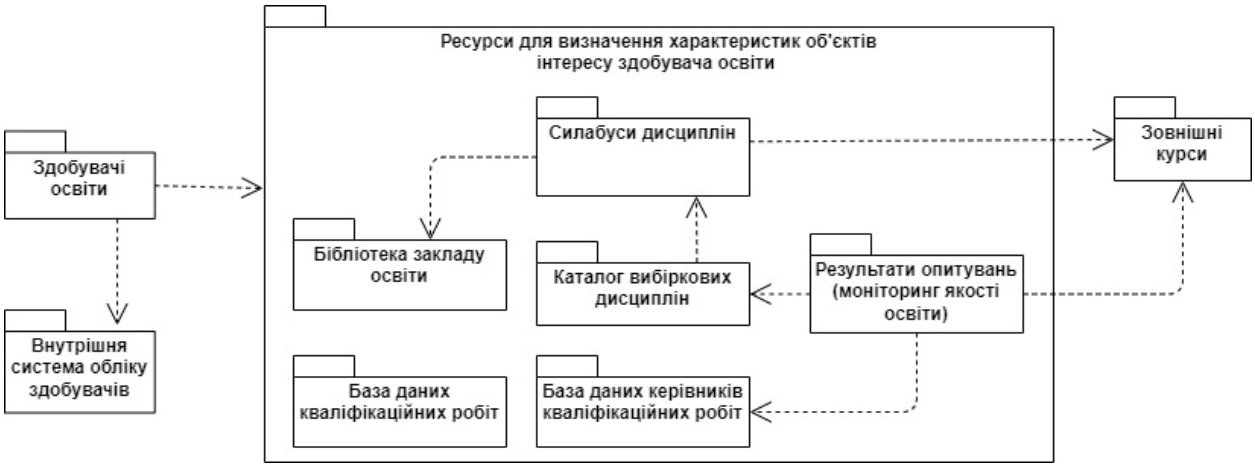


Рис.3.2. Ресурси для визначення характеристик об'єктів інтересу здобувачів та самих здобувачів в рамках формальної освіти

Розглянемо детальніше структуру, вміст та характеристики ресурсів, що можуть бути використані для побудови рекомендації відносно об’єктів інтересу здобувачів.

Силабуси дисциплін представляють собою документи, що містять короткі описи дисциплін. Формати та вміст силабусів унормовуються внутрішніми нормативними документами закладу освіти, але загалом вони повинні містити достатню інформацію для розуміння здобувачем цілей та задач дисципліни, політики та вимог курсу, основних результатів, які треба досягнути по закінченню курсу та іншу базову інформацію про дисципліну. Як правило, силабуси містять список рекомендованої літератури (в тому числі, посилання на ресурси з бібліотеки закладу освіти), а також, за необхідності, посилання (або рекомендації) на зовнішні курси, які можна пройти для

покращення результату цієї дисципліни чи для закріплення окремих обов'язкових чи додаткових навичок та інших цілей. Вміст силабусу представляє собою напівструктурований текст, який включає як метадані (наприклад, рівень вищої освіти, назва освітньої програми, мова навчання, контактна інформація викладача, пререквізити та постреквізити дисципліни, компетентності та програмні результати навчання тощо), так і довільні описи частин освітнього компоненту (наприклад, анотація дисципліни, опис форм та методів навчання, які використовуються під час занять, політики курсу тощо). Обсяги деталізованих силабусів можуть сягати до декількох тисяч слів, скорочений силабус може включати лише анотацію дисципліни та ключові метадані і мати обсяг 500-1000 слів.

Бібліотека закладу освіти використовується учасниками освітнього процесу для пошуку необхідної інформації за певною дисципліною, науковим напрямком чи сферою інтересів користувача. Технічно вона може представляти собою звичайний електронний каталог з назвами та анотаціями джерел або повнотекстову базу. Джерело з бібліотечного каталогу може саме виступати в якості об'єкту інтересу здобувача, або використовуватись для визначення характеристик інших об'єктів, наприклад, дисциплін. Представлення джерел в скороченому вигляді, як правило, передбачає використання метаданих для надання користувачеві тільки ключової інформації про джерело (наприклад, фрагмент каталогу електронної бібліотеки Державного університету інформаційно-комунікаційних технологій на рис.3.2).

Конкретний набір метаданих залежить від вимог закладу освіти чи освітньої платформи, але мінімальний набір може визначатись нижченаведеним списком:

– автор (або список авторів);

– мова документу;

– тип документу (посібник, підручник, методичні вказівки, конспект лекцій тощо);

– анотація.

Розмір документу може бути використаний як додатковий атрибут, оскільки при завантажені файлу користувач повинен розуміти, яку кількість пам’яті пристрою треба мати для завантаження.

Операційна ситема Linux: принципи роботи з файловою системою

Автор:	В.М. Черевик, Л.І. Танцюра, С.С. Коротков, В.О. Сосновий
Мова документу:	Українська
Розмір документу:	1.99 Мб
Тип документу:	Посібник

Посібник розкриває зміст сеансу роботи в Linux, виводить основні поняття терміналу і командного рядку, надає структуру файлової системи Linux, показує основні принципи роботи з файловою системою, має інформацію про процеси та доступи до об’єктів обробки, знайомить з основними прийомами роботи з текстовими даними та можливостями командної оболонки і описує базові відомості про текстові редактори vi і emacs. Ця основа знань і умінь надає можливість утворити базу на якій створюються умови якісно проводити подальші навчання в області операційних середовищ, систем і оболонок, а також, адміністрування системами.

Рис.3.2. Приклад скороченого представлення інформації про джерело електронної бібліотеки

Ще одним джерелом інформації для здобувача освіти є каталог вибіркових дисциплін – перелік дисциплін, які дозволяють сформувати так звану індивідуальну траєкторію здобувача. Ці дисципліни здобувач обирає самостійно за власними вподобаннями. Заклад освіти може при цьому накладати обмеження на вибір окремих компонент, наприклад, для забезпечення послідовності викладання взаємопов’язаних дисциплін. Дисципліни з цього каталогу виступають в якості об’єкту інтересу здобувача при формуванні його індивідуального навчального плану. Сам каталог може бути представлений деякою базою документів, що включають в себе силабуси або анотації навчальних дисциплін.

База даних кваліфікаційних робіт (репозиторій) може виступати в якості об’єкту інтересу здобувача з метою підбору тем робіт, що найбільше відповідають його інтересам або для визначення найбільш популярних серед здобувачів напрямків досліджень. Характеристики цього об’єкту інтересу з

точки зору рекомендаційної системи можуть включати метадані (назва галузі, назва спеціальності, рік виконання роботи, назва кафедри, дані керівника, перелік ключових слів) та повний або частковий опис контенту кваліфікаційної роботи. Окремі метадані, такі як ключові слова, напрямок досліджень, тематика, можуть бути визначені безпосередньо на основі аналізу вмісту текстів робіт чи з використанням тематичних індексів робіт.

База даних керівників кваліфікаційних робіт може виступати в якості об’єкту інтересу здобувача з метою підбору керівника, що найбільше відповідає тематиці досліджень здобувача або формування списку з таких керівників. Для формування рейтингових характеристик керівника можуть бути використані різні способи, наприклад, можна використовувати результати опитувань здобувачів, дані рейтингування викладачів на рівні закладу освіти по різним видам робіт, досвід викладача в керуванні кваліфікаційними роботами за тематикою тощо.

Для отримання числових рейтингових оцінок об’єктів інтересу в рамках закладу освіти можна використовувати статистичні звіти, що містять результати опитувань здобувачів, викладачів та інших стейкхолдерів освітнього процесу, в тому числі, рейтингування викладачів, рейтингування дисциплін чи окремих видів навчальних робіт тощо. При цьому для побудови рейтингів можуть використовуватись шкали оцінювання, розглянуті в розділі 2.

Як було зазначено раніше, зовнішні курси не є частиною внутрішніх ресурсів закладу освіти, що використовуються в освітньому процесі, і, враховуючи надзвичайну різноманітність подібних курсів, не представляється можливим інтегрувати їх, як частину рекомендаційної системи. Тим не менш, інформація про взаємодію з такими курсами можна отримати на основі опитувань здобувачів (наприклад, анкетування наприкінці курсу з метою виявлення якими саме курсами здобувачі скористались і як вони їх оцінюють), а також на основі інформації з силабусів навчальних дисциплін, де окремі зовнішні курси можуть бути вказані в якості рекомендації викладачів.

3.1.3 Визначення характеристик користувачів рекомендаційної системи

Основними користувачами рекомендаційної системи в сфері освіти є здобувачі (студенти). Інформація безпосередньо про здобувачів, як потенційних користувачів рекомендаційної системи, може бути отримана з баз даних внутрішньої системи обліку здобувачів. Окрім інформації про персональні, демографічні, соціальні та інші характеристики здобувача, для побудови «портрету» користувача можуть використовуватись додаткові дані, наприклад про його успішність, участь в позанавчальних заходах (конференції, конкурси тощо). Зокрема, актуальною для рекомендаційної системи закладу освіти може бути інформація з індивідуальних навчальних планів студентів, що включає дані про вивчені дисципліни та отримані оцінки, в тому числі, вибіркові компоненти, які студент вже обрав для вивчення в попередніх навчальних семестрах.

Формальна модель здобувача освіти, як користувача рекомендаційної системи, може включати різні аспекти і описувати його інтереси та взаємодію з освітніми ресурсами. Представимо i-го здобувача у вигляді кортежу U_i, що описується виразом (3.1):

$$U_i =< ep, md, cd, w >_i, i = \overline{1, N}, \tag{3.1}$$

де N – кількість здобувачів, інформація про яких є в системі; ep – освітня програма, на якій навчається здобувач; md – інформація про обов’язкові дисципліни індивідуального навчального плану здобувача; cd – інформація про вибіркові дисципліни індивідуального навчального плану здобувача; w – множина текстових пошукових запитів здобувача. В умовах «холодного старту» наявною є тільки інформація про освітню програму, на якій навчається студент. Всі інші дані з’являються в результаті виконання студентом індивідуального навчального плану та взаємодії з інформаційними ресурсами закладу освіти.

Освітня програма ep є елементом вектору всіх освітніх програм закладу освіти $ep \in EP$, загальна кількість освітніх програм визначається величиною NP. При цьому, кожна освітня програма може бути описана кортежем (3.2):

$$ep =< EPname, EPkeywords, EPdescription >, \quad (3.2)$$

де $EPname$– назва освітньої програми; $EPkeywords$– множина ключових слів освітньої програми; $EPdescription$ – текстовий опис освітньої програми. Близькість освітніх програм між собою може бути визначена на основі відстані між ембедінгами назв, ключових слів та/або текстових описів освітніх програм.

Кожна величина md інформації про обов’язкові дисципліни індивідуального навчального плану студента може бути представлена множиною пар (3.3):

$$md = \{< Mdiscipline, Mmark >\}, \quad (3.3)$$

де $Mdiscipline$ – обов'язкова дисципліна, яку вивчав здобувач; $Mmark$ - оцінка за 100-бальною шкалою, яку отримав здобувач з обов’язкової дисципліни. Оцінку здобувача з дисципліни можна в якомусь сенсі інтерпретувати, як ступінь його зацікавленості дисципліною під час навчального процесу. Якщо здобувач тільки почав навчання (1 семестр 1 курсу), то множина (3.4) є пустою. Після 1 семестру навчання множина вже обов’язково буде містити дані з індивідуального навчального плану студента, що зменшує проблему «холодного старту». Елемент $Mdiscipline$ є складовою навчального плану за певною освітньою програмою.

Інформація про вибіркові дисципліни індивідуального навчального плану cd здобувача аналогічно до md може бути представлена множиною пар (3.4):

$$cd = \{< Cdiscipline, Cmark >\}, \quad (3.4)$$

де $Cdiscipline$ – вибіркова дисципліна, яку вивчав здобувач; $Cmark$ - оцінка за 100-бальною шкалою, яку отримав здобувач з вибіркової дисципліни. На відміну від обов'язкових дисциплін, наявність у множині cd деякої дисципліни вже можна інтерпретувати, як виявлення інтересу здобувача, а високі оцінки з такої дисципліни можуть використовуватись для посилення загального рейтингу дисципліни. Якщо здобувач ще не вивчав жодної вибіркової дисципліни, то множина (3.4) відповідно є пустою.

Множина w визначає текстові пошукові запити здобувача і фактично описує область його інтересів. В загальному випадку ця множина може представляти собою просто набір деяких текстових фрагментів, але для побудови більш точних рекомендацій можна розділити всю множину текстових запитів w на підмножини (3.5):

$$w = \{wCD, wL, wC, wT\}, \tag{3.5}$$

де wCD – множина пошукових текстових запитів, які користувач задавав на сторінці каталогу вибіркових дисциплін; wL – множина пошукових текстових запитів, які користувач задавав на сторінці електронної бібліотеки; wC – множина пошукових текстових запитів, які користувач задавав на сторінці пошуку керівника кваліфікаційної, наукової чи іншої роботи; wT - множина пошукових текстових запитів, які користувач задавав на сторінці репозиторію кваліфікаційних робіт. Зазначені в (3.5) множини можуть бути використані як для побудови конкретної рекомендації, що стосується цілі пошуку, так і для побудови будь-якої іншої рекомендації, оскільки сукупно ці запити можна інтерпретувати як сферу інтересів здобувача.

3.2 Моделювання структури рекомендаційної системи

Для формування рекомендацій відносно визначених для освітньої сфери об'єктів інтересу можуть використовуватись:

– результати цільових опитувань/оцінювань відносно об’єктів інтересу (отримання явної оцінки користувачів);

– результати аналізу інформації про поведінку користувачів відносно об’єкту інтересу та їх взаємодії (отримання неявної оцінки користувачів);

– результати аналізу контенту об’єкту інтересу з метою визначення рівня релевантності інтересів користувача та характеристик об’єкту інтересу.

Структура рекомендаційної системи в освітній сфері для випадку вищої освіти або освіти дорослих може бути представлена моделлю на рис.3.3.

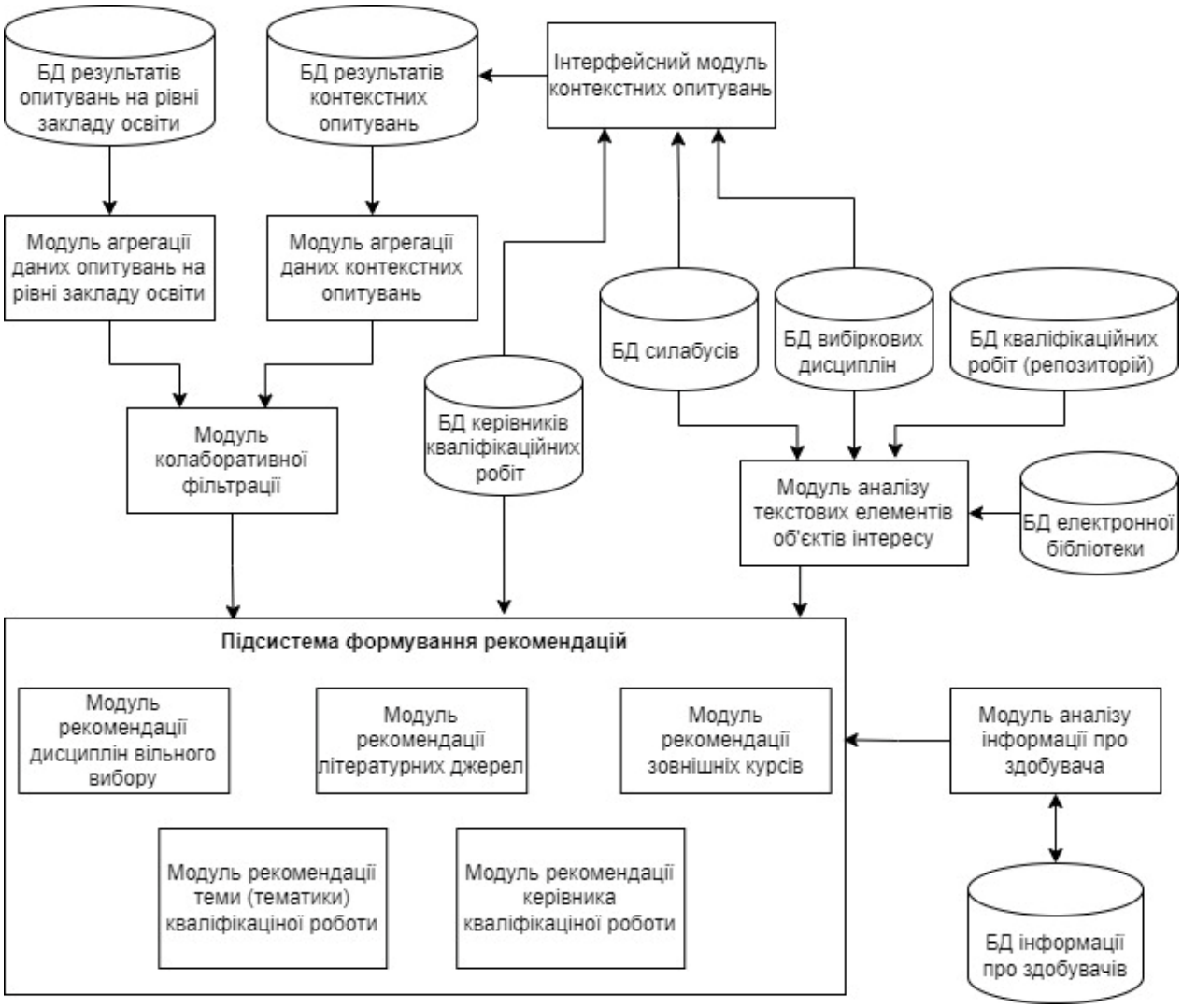


Рис.3.3. Структурна модель рекомендаційної системи в сфері освіти

В структурі системи передбачено виділення окремого модуля для формування рекомендації по кожному об’єкту інтересу окремо, оскільки

деталізовані моделі побудови рекомендації для кожного з них відрізняються об’єктами інтересу та, відповідно, базами даних, які вони використовують в своїй роботі. Для реалізації рекомендаційної системи пропонується використовувати гібридний підхід.

Частину інформації, що включає в себе результати опитувань здобувачів, можна, в тому числі, обробляти з використанням колаборативного підходу. При цьому, для формування рекомендацій керівника кваліфікаційної роботи, дисципліни вільного вибору та зовнішніх курсів можна очікувати низький рівень розрідженості матриці взаємодії з двох причин:

– опитування здобувачів є обов’язковою частиною освітнього процесу і до них залучається великий відсоток здобувачів;

– кількість об’єктів інтересу в зазначених класах є невеликою для кожного окремого модуля побудови рекомендацій і матриця взаємодій може розглядатися частково в контексті деякої освітньої програми, спеціальності чи іншого параметру відбору.

Контекстні опитування є додатковим засобом для визначення ставлення користувача до об’єктів інтересу і можуть бути отримані поза межами основної системи обов’язкових опитувань.

Що стосується модулів рекомендацій теми кваліфікаційної роботи та літературних джерел, то, зважаючи на те, що метадані про об’єкти інтересу (назва, рік видання, автор тощо) та вміст самих об’єктів представлено текстовими даними, пропонується для їх аналізу застосовувати методи обробки текстів природною мовою (Natural Language Processing, Text Mining). Також модуль аналізу текстових елементів доцільно застосовувати для формування окремих характеристик силабусів та вибіркових дисциплін (наприклад, ключових слів, визначень, тематики та інших), які в подальшому можна використовувати для визначення схожості з інтересами здобувача.

3.3 Розробка моделі даних рекомендаційної системи для сфери освіти на основі нечіткої логіки

Оскільки студенти регулярно приймають участь у певних заходах щодо оцінювання тих чи інших навчальних ресурсів, дисциплін, літературних джерел, викладачів, то цю інформацію можна використовувати для побудови рекомендації. Типовим способом оцінювання є використання числових шкал. Мінімізація залежності впливу діапазону та формату шкали оцінювання на процес побудови рекомендації в освітній сфері потребує створення моделі, яка дозволить легкий перехід між різними шкалами оцінювання параметрів, що використовуються при формуванні рекомендації, та забезпечить при цьому збереження семантичної складової оцінки.

Для побудови моделі представлення даних рекомендаційної системи в сфері освіти користувачів рекомендаційної системи, що мають потребу в консультації/рекомендації, можна ввести додаткове обмеження щодо здобувачів, що вони відносяться до певної освітньої програми.

Для опису моделей даних рекомендаційної системи пропонується застосувати нечітку логіку. Для даної задачі це обумовлюється наступними причинами:

1. Лінгвістичне представлення є більш природним з точки зору сприйняття людини у порівнянні з числовим представленням – нечітка логіка в цьому контексті дозволяє співставити числовим значенням текстові визначення.

2. Вибір шкал для оцінювальних параметрів об’єктів інтересу, характеристик користувачів чи результатів взаємодії користувачів з об’єктами інтересу не завжди є однозначним – будь-яка, навіть зовсім незначна зміна числової шкали оцінювання призводить до іншого абсолютного числового значення оцінки, що потребує внесення змін в алгоритм формування рекомендації.

3. Нечіткі значення можуть бути використані для випадків, коли параметри, що оцінюються, не мають заздалегідь визначеного значення або

обчислення цього значення ускладнено – це стосується в першу чергу якісних характеристик об’єктів інтересу та користувачів.

Розглянемо порядок побудови нечіткої моделі даних, що представляють собою деяку ранжовану оцінку, на прикладі показника рейтингу дисципліни.

Рейтинг дисципліни може бути сформований за результатами опитувань студентів. Опишемо цей рейтинг як лінгвістичну змінну D_rating. Її терм-множину представимо виразом (3.6), що включає значення, визначені відповідно до психометричної шкали Лайкерта на 5 градацій:

$$T_{D_{rating}} = \{\text{"дуже низький", "низький", "середній", "вище середнього", "високий"}\}. \quad (3.6)$$

Еквівалентні назви термів англійською мовою для забезпечення комп’ютерного моделювання можуть визначатися парами: “дуже низький” – “very low”, “низький” – “low”, “середній” – “normal”, “вище середнього” – “above normal”, “високий” – “high”, тоді терм множина (3.6) матиме вигляд (3.7)

$$T_{D_rating} = \{\text{"very low", "low", "normal", "above normal", "high"}\}. \quad (3.7)$$

Для визначення рейтингу можна використовувати альтернативні терм-множини, які визначають відтінки відношення студента до дисципліни (табл.3.1). Слід зауважити, що різні представлення термів терм-множини при цьому не впливають на потужність самої множини та, в загальному випадку, на спосіб обчислення нечіткого значення – для всіх альтернатив терму можна обрати одну й ту саму функцію належності.

Таблиця 3.1. Альтернативні терм-множини для визначення рейтингу дисципліни

Відтінок відношення	**Терм 1**	**Терм 2**	**Терм 3**	**Терм 4**	**Терм 5**
Ставлення	Дуже погано	Погано	Ні добре, ні погано (нейтральне ставлення)	Добре	Дуже добре
Важливість	Абсолютно не важливо	Неважливо	Важко відповісти (нейтрально)	Важливо	Дуже важливо
Задоволеність	Дуже незадово-лений	Незадово-лений	Нейтральний	Задоволений	Дуже задоволений
Намір обрати дисципліну	Точно не оберу	Швидше за все не оберу	Рівна ймовірність (важко відповісти)	Швидше за все оберу	Точно оберу

Універсальна множина E для лінгвістичної змінної D_rating визначається в кожному конкретному випадку в залежності від способу обчислення числового значення рейтингової оцінки. Наприклад, якщо рейтинг розраховується на основі однієї рангової оцінки в діапазоні [1; 5], то універсум E=[1; 5], а функції належності відповідних термів можуть бути визначені, наприклад, трикутними функціями виду (3.8-3.12) з графіками функцій належності, представленими на рис.3.4. Перший та останній терми для терм-множини (3.6) («дуже низький» та «високий») мають вигляд усічених трикутників.

$$\mu_{verylow}(x) = \begin{cases} 2 - x, \text{ при } 1 < x \leq 2 \\ 0, \text{ інакше} \end{cases}; \tag{3.8}$$

$$\mu_{low}(x) = \begin{cases} x - 1, \text{при } 1 < x \leq 2 \\ 3 - x, \text{ при } 2 < x \leq 3; \\ 0, \text{ інакше} \end{cases} \tag{3.9}$$

$$\mu_{normal}(x) = \begin{cases} x - 2, \text{при } 2 < x \leq 3 \\ 4 - x, \text{ при } 3 < x \leq 4; \\ 0, \text{ інакше} \end{cases} \tag{3.10}$$

$$\mu_{abovenormal}(x) = \begin{cases} x - 3, \text{при } 3 < x \leq 4 \\ 5 - x, \text{ при } 4 < x \leq 5; \\ 0, \text{ інакше} \end{cases} \tag{3.11}$$

$$\mu_{high}(x) = \begin{cases} x - 4, \text{ при } 4 < x \leq 5 \\ 0, \text{ інакше} \end{cases}. \tag{3.12}$$

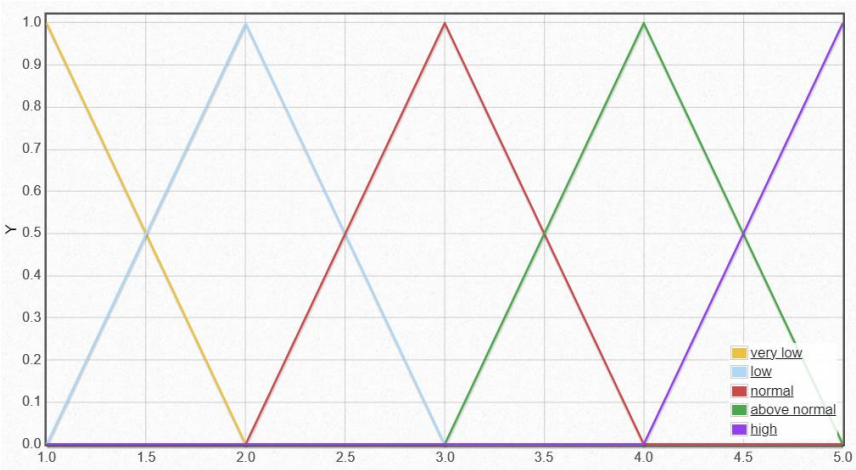


Рис. 3.4. Графіки функцій належності термів лінгвістичної змінної D_rating при використанні 5 градацій шкали Лайкерта для ранжування рейтингу дисципліни

Параметри трикутних функцій для лінгвістичної змінної D_rating напряму залежать від кількості градацій шкали. В загальному випадку, при використанні іншої дискретної шкали з кроком градації в 1 пункт шкали, наприклад, шкали Лайкерта з 7 градаціями, функція належності терма $x \in X_{D_rating}$ може бути визначена загальними виразами (3.13-3.15):

$$\mu(x,a)=\begin{cases}x-a+1, \text{при } a-1<x\le a\\ a+1-x,\ \text{при } a<x\le a+1\\ 0,\ \text{інакше}\end{cases}\Bigg|\ \forall a\in[dr_{min}+1, dr_{max}-1]; \quad (3.13)$$

$$\mu(x,a)=\begin{cases}dr_{min}+1-x,\ \text{при } dr_{min}<x\le dr_{min}+1\\ 0,\ \text{інакше}\end{cases}\Bigg|\ a=dr_{min}; \quad (3.14)$$

$$\mu(x,a)=\begin{cases}x-dr_{max}-1,\ \text{при } dr_{max}-1<x\le dr_{max}\\ 0,\ \text{інакше}\end{cases}\Bigg|\ a=dr_{max}; \quad (3.15)$$

де dr_{min} визначає числове значення крайньої лівої градації шкали, dr_{max} – числове значення крайньої правої градації шкали. В такому випадку універсум визначається виразом (3.16):

$$E=[dr_{min}, dr_{max}]. \quad (3.16)$$

Для шкали Лайкерта будь-якої градації dr_{min} як правило становить 1, але в деяких випадках шкала може бути розширена значенням 0.

Особливістю процесу формування рейтингової оцінки є те, що в окремих випадках вона може описуватись багатокритеріальною моделлю – рейтингова оцінка визначається по декільком показникам, які включають, наприклад, кваліфікацію викладача, відповідність дисципліни силабусу (документ, що містить опис самого курсу, його політик та іншу інформацію, необхідну для ознайомлення з дисципліною), сучасність навчальних матеріалів, відповідність методів навчання, цікавість навчального матеріалу, складність навчального матеріалу, прозорість критеріїв оцінювання, об’єктивність оцінювання (рис.3.5).

Оцінку рівня кожного показника можна представити числовими значеннями за шкалою від 1 до 5 та окремим значенням «Важко відповісти», яке може означати невпевненість студента в своїй оцінці, небажання надавати оцінку чи відсутність досвіду взаємодії (студент не відвідував заняття). Таким чином, терм «Важко відповісти» можна співвіднести з числовим значенням 0.

Оцініть викладання дисципліни "DevOps" *

	5	4	3	2	1	Важко відповісти
Кваліфікація викладача	○	○	○	○	○	○
Відповідність дисципліни силабусу	○	○	○	○	○	○
Сучасність навчальних матеріалів	○	○	○	○	○	○
Відповідність методів навчання	○	○	○	○	○	○
Цікавість навчального матеріалу	○	○	○	○	○	○
Складність навчального матеріалу	○	○	○	○	○	○
Прозорість критеріїв оцінювання	○	○	○	○	○	○
Об'єктивність оцінювання	○	○	○	○	○	○

Рис. 3.5. Приклад параметрів опитування для формування рейтингової оцінки дисципліни

Для ситуацій, коли рейтингова оцінка формується на основі декількох критеріїв визначимо додатково деяку функцію (3.17), що дозволяє розрахувати значення рейтингу:

$$f(mark_1, mark_2, \dots, mark_n), \tag{3.17}$$

де $mark_i$ – значення показника, що використовується при обчисленні рейтингу, n – кількість показників. Слід зазначити, що у випадку використання дискретних шкал для визначення значення показника, кількість градацій в шкалах в загальному випадку може відрізнятись.

При роботі з показниками, що визначають деяке відношення типу «відповідність» (дисципліна відповідає інтересам студента, тематика досліджень викладача відповідає напрямку кваліфікаційної роботи чи

науковим інтересам студента і т.п.) пропонується використовувати один з двох підходів:

1. Обчислювати відповідність, оцінюючи її так само, як дискретну чи безперервну рейтингову величину в універсумі $E = [dr_{min};\ dr_{max}]$. Терм-множина нечіткої змінної відповідності в цьому випадку може бути описана як *T*={"абсолютно не відповідає", "відповідає частково", "більш-менш відповідає", "відповідає значною мірою", "повністю відповідає"}. Функції належності відповідних термів можуть бути описані як трикутні, гаусові чи інші.

2. Обчислювати відповідність як ступінь істинності – тобто як безперервну величину в універсумі $E = [0; 1]$. В якості функції належності для цього випадку може бути використано, наприклад, Гаусову функцію (3.18):

$$\mu(x) = exp\left[-\left(\frac{x-c}{\delta}\right)^2\right], \tag{3.18}$$

де c – центр нечіткої множини, δ – крутизна функції належності.

Приклад нечіткої змінної, що визначає показник відповідності з терм множиною *T*={"match","does not match"} і Гаусовими функціями належності $\mu_{match}(x) = exp\left[-\left(\frac{x-0.7}{0.2}\right)^2\right]$ та $\mu_{does\ not\ match}(x) = exp\left[-\left(\frac{x-0.3}{0.2}\right)^2\right]$ наведено на рис.3.6.

Побудуємо рекомендацію для вибору дисципліни, використовуючи вихідну нечітку змінну $DisciplineRecommendation$, що описується терм множиною $T_{DisciplineRecommendation} = \{"recommended",$ $"not\ recommended"\}$ та гаусовими функціями належності $\mu_{recommended}(x) = exp\left[-\left(\frac{x-10}{5}\right)^2\right]$ та $\mu_{not\ recommended}(x) = exp\left[-\left(\frac{x}{3}\right)^2\right]$. Універсум нечіткої змінної для цього прикладу визначимо на діапазоні $E = [0;\ 10]$ (діапазон універсуму визначається конкретними вимогами рекомендаційної системи). Наведені функції належності утворюють невелике зміщення області позитивного рішення щодо рекомендації вліво від центру

діапазону допустимих значень, що дозволяє зменшити порогове значення для позитивної рекомендації. Графічне представленні нечіткої вихідної змінної наведено на рис.3.7.

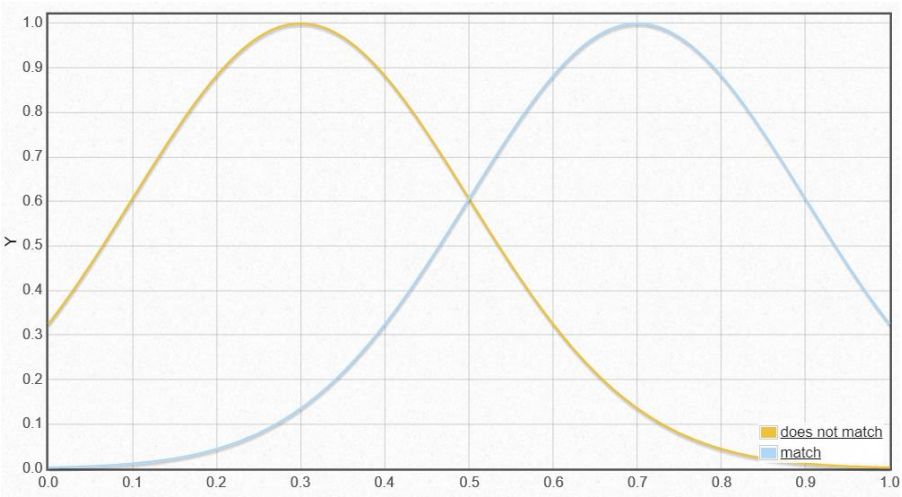


Рис.3.6. Приклад відображення нечіткої змінної, що визначає показник відповідності з терм множиною T={"match","does not match"} (функція належності – гаусова)

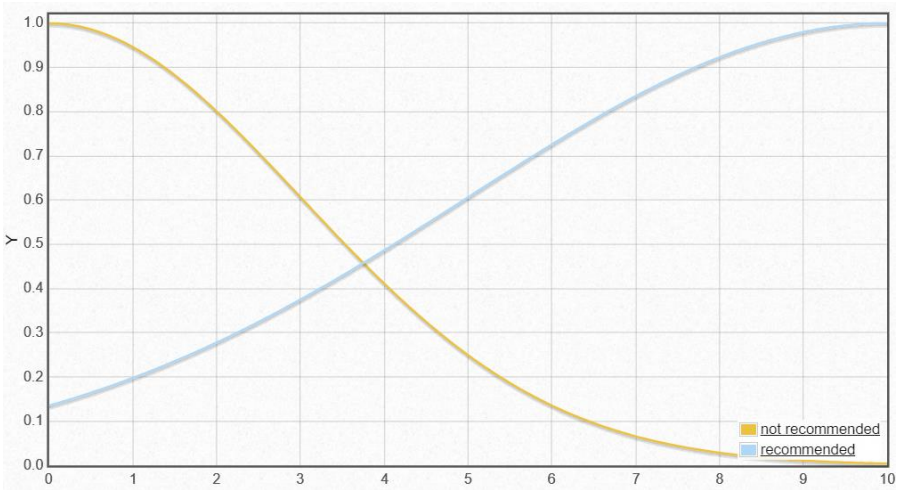


Рис. 3.7. Приклад відображення вихідної нечіткої змінної для формування рекомендації

Для отримання рекомендації необхідно побудувати правила нечіткого виведення, які описують залежність нечіткої вихідної змінної від нечітких вхідних змінних. Правила мають продукційну форму “if–then”.

Приклади правил, які можуть бути використані для побудови рекомендації, та результати моделювання нечіткого виведення з антецедентами рейтингу дисципліни 1.7, рівня відповідності дисципліни інтересам студента 0.2 та рівня відповідності пререквізитів дисципліні 0.7 на основі алгоритму Мамдані наведено на рис.3.8.

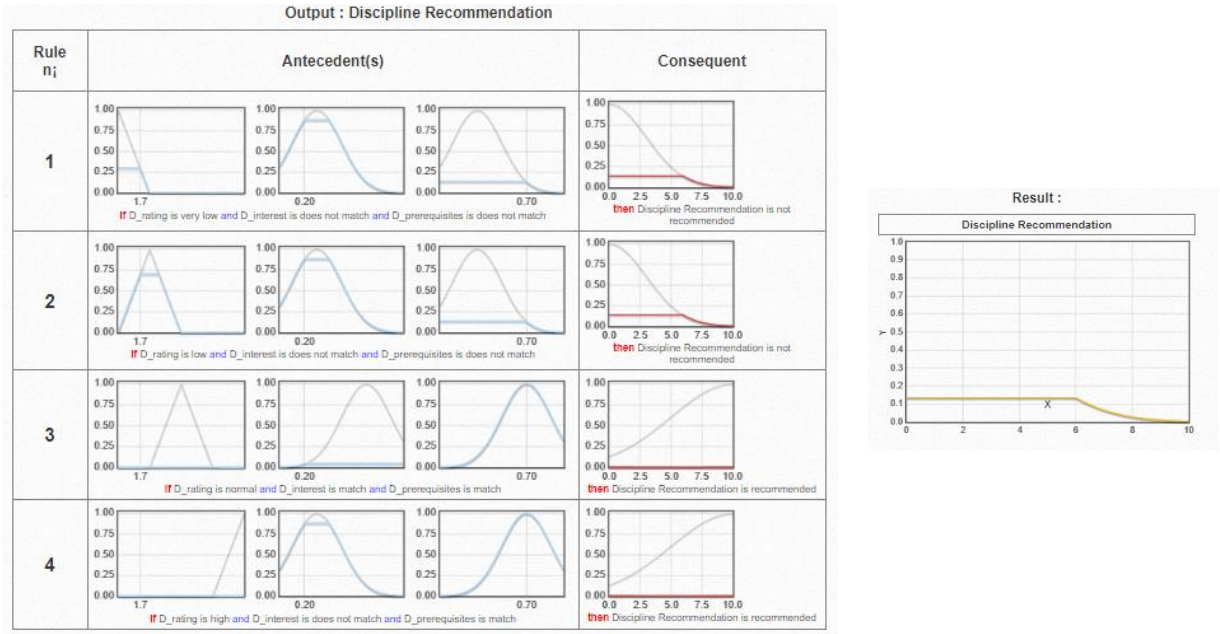


Рис. 3.8. Приклад результатів моделювання нечіткого виведення при формуванні рекомендації дисципліни

Запропонований підхід є універсальним в межах контексту предметної галузі сфери освіти і може бути розширений на інші показники рекомендаційної системи. В таблиці 3.2 представлено види показників, які можуть бути використані для побудови описаних раніше рекомендацій, і значення яких можуть бути перетворені в нечіткі змінні відповідно до наведених моделей.

Таблиця 3.2 – Види показників, що можуть бути використані для побудови рекомендацій та приведені до нечіткого формату

Об'єкт рекомендації	**Показник**	**Спосіб визначення показника**
Дисципліна	Рейтинг дисципліни	Число в деякому діапазоні
	Відповідність дисципліни інтересам студента	Число в деякому діапазоні або логічна величина «істина (відповідає)/не істина (не відповідає)»
	Відповідність наявної освіти студента пререквізитам дисципліни	Число в деякому діапазоні або логічна величина «істина/не істина»
Джерело з бібліотеки закладу освіти	Рейтинг джерела (наскільки його рекомендують інші учасники освітнього процесу)	Число в деякому діапазоні
	Частота використання джерела	Число в деякому діапазоні
	Відповідність джерела інтересам студента	Число в деякому діапазоні або логічна величина «істина (відповідає)/не істина (не відповідає)»
Тема (або тематика дослідження) кваліфікаційної роботи чи наукової роботи, проєкту тощо	Рівень популярності тематики/напрямку	Число в деякому діапазоні
	Відповідність тематики/напрямку інтересам студента	Число в деякому діапазоні або логічна величина «істина (відповідає)/не істина (не відповідає)»
Потенційний керівник кваліфікаційної роботи, наукової	Рейтинг викладача	Число в деякому діапазоні
	Відповідність спеціалізації викладача тематиці майбутніх досліджень	Число в деякому діапазоні або логічна величина «істина (відповідає)/не істина (не відповідає)»

роботи, проєкту тощо		
Зовнішній курс (*за наявності відповідної інформації в базі рекомендаційної системи)	Рейтинг курсу	Число в деякому діапазоні
	Частота використання курсу	Число в деякому діапазоні

3.4 Розробка методу формування рекомендацій по об’єктам інтересу з використанням гібридного підходу

Формування рекомендації пропонується здійснювати з використанням гібридного підходу що включає різні методи роботи з наявними даними, зокрема, обробку оцінок користувачів, отриманих на основі опитувань, визначення близькості характеристик об’єктів інтересу та сфери інтересів користувачів, колаборативну фільтрацію для врахування їх взаємодії користувачів з об’єктами інтересу.

Етапи методу автоматизованого формування рекомендації на основі гібридного підходу можна описати алгоритмом на рис.3.9. Для побудови рекомендації на першому кроці визначається об’єкт інтересу користувача. Інформація про кожен з об’єктів інтересу представляється такими параметрами, які дозволили б реалізацію наступних кроків методу.

Для розрахунку ступеня релевантності об’єкту до інтересів користувача можуть використовуватись різні метрики подібності. Також, сама релевантність може трактуватись різними способами: це може бути обчислення семантичної подібності (наприклад, з використанням методів машинного навчання) або відстань між векторизованими представленнями об’єкта та інформації про інтереси користувача (наприклад, косинусна відстань між вектором пошукового запиту користувача та вектором об’єкту інтересу).

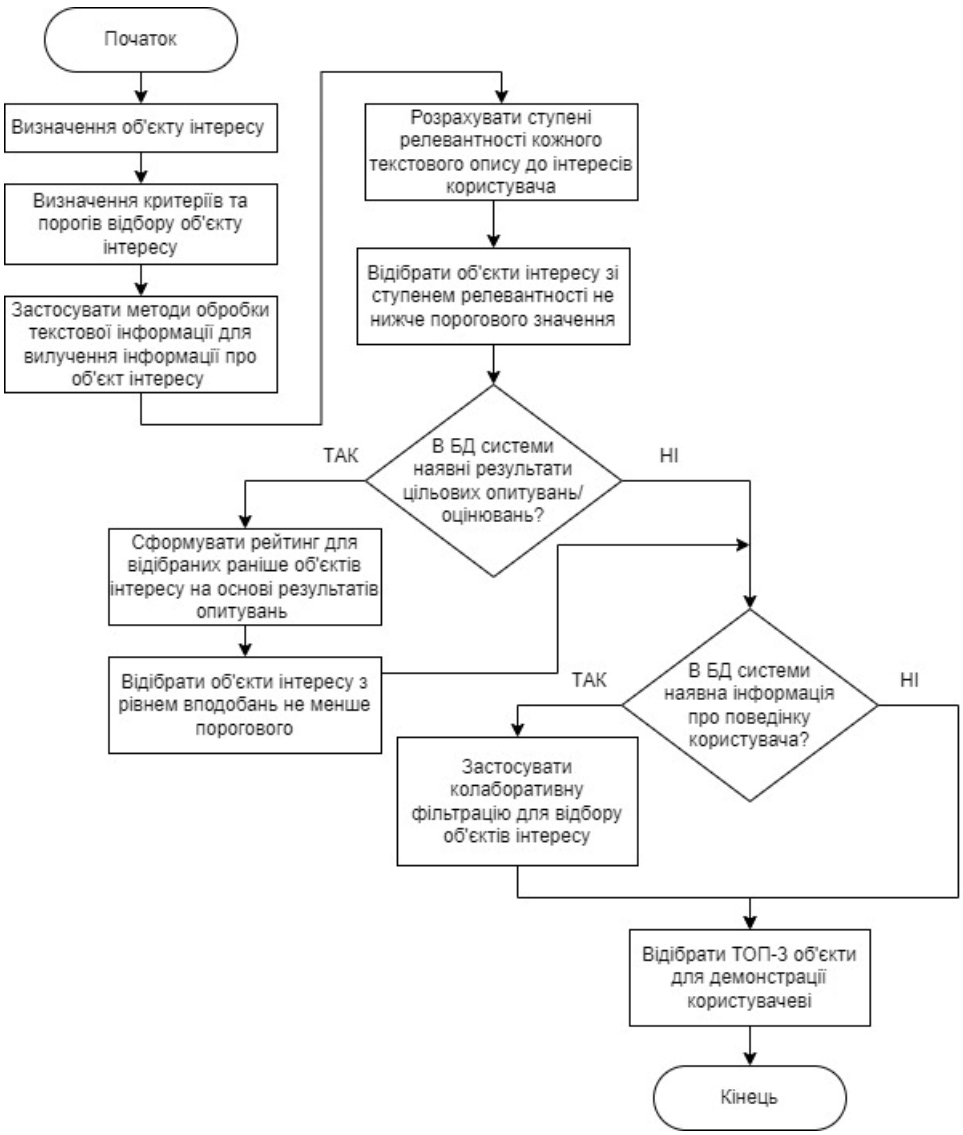


Рис.3.9. Етапи методу автоматизованого формування рекомендації на основі гібридного підходу

Після етапу попереднього відбору об’єктів інтересу на основі їх контенту система повинна використовувати дані взаємодії користувачів з об’єктами. Для освітньої сфери це можуть бути взаємодії двох видів:

– явні у вигляді цільових опитувань користувачів про об’єкти інтересу (оцінка студентами якості викладання дисциплін, рейтинги викладачів, оцінки підручників, посібників тощо);

– неявні, які отримуються за результатами моніторингу дій користувачів в системі, що містить об’єкти інтересу (наприклад, відвідування сторінки дисципліни в системі дистанційного навчання, відвідування сторінки підручника на сайті бібліотеки і т.п.).

Використання даних цільових опитувань дозволяє сформувати попередню користувацьку оцінку для переважної більшості об’єктів, оскільки такі цільові опитування є обов’язковим елементом моніторингу якості освітнього процесу. Виключенням можуть бути такі об’єкти інтересу, які з’являються в освітньому процесі вперше (новий викладач, нова дисципліна, новий підручник, новий напрямок досліджень тощо). Однак, завдяки етапу відбору на основі аналізу вмісту текстового контенту для цих об’єктів можна отримати деяку рейтингову оцінку, співставивши їх зі схожими об’єктами відповідно до обраної метрики схожості.

Особливістю об’єктів інтересу здобувача є те, що для кожної групи наявний короткий чи повний текстовий опис, що характеризує об’єкт. Для побудови рекомендації вимоги до такого текстового опису визначаються наборами даних, які можна сформувати в умовах реального закладу освіти:

– для дисциплін вільного вибору здобувача мінімальний контент повинен складатись з назви та анотації дисципліни, в ідеальному варіанті – повного опису вмісту дисципліни у формі робочої програми чи силабусу;

– інформаційні джерела можуть бути представлені безпосередньо текстовим контентом або метаінформацією про них (автори, ключові слова тощо);

– дані про керівників повинні містити текстові описи наукових та професійних інтересів викладача та/або список публікацій, проєктів тощо;

– курсові, дипломні роботи, творчі проєкти можуть бути представлені короткими описами, що включають лише назву роботи, але бажаним варіантом представлення, що дозволить отримати більш точну рекомендацію, є наявність анотації чи повного текстового опису роботи/проєкту.

Критерії відбору та порогові значення для подальшого відбору можуть коригуватись в залежності від групи об’єктів або користувацьких налаштувань. Оскільки кожен з визначених об’єктів інтересу містить деякий текстовий опис, на першому етапі формування рекомендації застосовуються методи обробки текстової інформації (Natural Language Processing) для вилучення даних про об’єкти інтересу. Перелік даних, що вилучаються, залежить від типу об’єкту. Слід зазначити, що для зменшення часу формування рекомендації, ключова інформація про об’єкти інтересу, що використовується в подальших етапах методу, може бути сформована заздалегідь та оновлюватись тільки у випадках зміни контенту об’єктів інтересу чи додаванні/видаленні об’єкту.

3.5 Алгоритм застосування методики побудови рекомендаційної системи для сфери освіти на основі гібридного підходу

Побудова рекомендаційної системи на основі розробленої методики передбачає виконання наступних кроків за наступним алгоритмом:

1) визначення наявних ресурсів, що містять інформацію про об’єкти інтересу, форматів представлення даних, особливостей доступу до інформації;

2) визначення інформації про особливості обчислення рейтингу об’єкту інтересу – шкала оцінювання, спосіб оцінювання, рівень залученості та періодичність залучення користувачів в процес оцінювання;

3) визначення інформації про користувачів, важливої для побудови рекомендації – види інформації, ресурси, що містять зазначену інформацію та особливості доступу до них;

4) визначення переліку нечітких даних системи та параметрів функцій належності нечітких змінних;

5) визначення необхідних методів попередньої обробки для текстових елементів даних;

6) визначення метрик подібності текстових елементів даних;

7) побудова правил відбору об’єктів рекомендацій на кожному з проміжних етапів відбору.

Результати 1, 2, 3 кроків дозволяють побудувати архітектуру майбутньої системи, визначити складові елементи необхідних баз даних, наявність чи відсутність окремих модулів системи в залежності від характеру інформації. Архітектура визначає структуру системи, вказуючи на компоненти, їхні взаємодії та взаємозалежності. Наприклад, у випадку відсутності даних про цільові опитування студентів, модулі, які відповідають за агрегацію цих даних та їх подальшу обробку, будуть також відсутні в системі. Якщо аналіз текстової складової інформації про об’єкти інтересу та самого користувача показує, що вона обмежується дуже короткими текстами або взагалі окремими словами чи словосполученнями довжиною не більше 3 слова, то модуль попередньої обробки текстів може бути вилучений з архітектури майбутньої системи або реалізовувати тільки етап токенізації, а задача визначення релевантності інформації про користувача чи його пошукового запиту до характеристик об’єкту інтересу може бути зведена до визначення подібності окремих слів. У випадку наявності повноцінних текстових фрагментів як в інформації про об’єкт інтересу, так і в інформації про користувача, система обов’язково повинна містити модулі, що забезпечують необхідну попередню обробку з урахуванням якості та структури текстової інформації (крок 5).

4-й крок передбачає визначення переліку нечітких даних системи та параметрів функцій належності нечітких змінних, що вимагає від розробника системи розділити наявні дані на чіткі та ті, що мають деякі ознаки недосконалості чи представлені якісними значеннями. В п. 3.3 наведено типові види даних освітньої сфери, що можуть мати представлення у вигляді нечітких значень, однак який саме формат та спосіб представлення даних використовувати в реальній рекомендаційній системі, залежить від технічних, організаційних та інших особливостей її реалізації. Наприклад, якщо в системі вже є якісні оцінки, то переведення їх у формат нечітких даних може бути виконано за наступною схемою:

– визначення переліку лінгвістичних значень величини;

– визначення числових діапазонів для кожного лінгвістичного значення;

– вибір функції належності для кожного лінгвістичного значення.

Слід зазначити, що специфікою освітньої сфери є велика кількість саме якісних, а не кількісних оцінок. При зміні градацій чи діапазону шкали оцінювання розробник системи повинен внести зміни лише в опис лінгвістичних змінних, а продукційні правила для формування рекомендації при цьому не змінюються, що підвищує гнучкість системі в контексті мінливості вимог.

5 та 6 кроки передбачають аналіз специфіки наявної текстової інформації. Для якісної реалізації цих кроків розробник рекомендаційної системи повинен врахувати наступні аспекти:

– розміри вхідної текстової інформації (кількість слів в одному наборі даних) – тут важливим фактором є не просто розмір окремого набору даних, а більше взаємне співвідношення розміру текстового опису одного елементу по відношенню до іншого; наприклад, якщо пошуковий запит користувача містить 1-2 слова, то порівнювати векторне представлення такого запиту з вектором анотації з 200-500 слів недоцільно, а треба використовувати методи, що ґрунтуються на пошуку подібності за ключовими словами;

– синтаксична, граматична, лексична якість тексту – тексти низької якості містять різні види шуму (помилки друку, спеціальні символи, абревіатури, скорочення, жаргонізми тощо) і потребують реалізацію додаткових специфічних функцій попередньої обробки;

– семантична якість тексту – якщо в тексті відсутні токени, що дозволяють визначити семантичну близькість до іншого тексту, то семантична цінність такого тексту є низькою, прикладом такого тексту може виступати пошуковий запит, що містить загальновживані слова та/або стоп-слова (займенники, прийменники, частки тощо), наприклад пошуковий запит

«корисна література» не містить жодного семантично значущого токену, який можна було б використати для порівняння з описом дисципліни.

Загальні рекомендації щодо використання тих чи інших методів попередньої обробки зведені в таблиці 3.3.

Таблиця 3.3 – Рекомендації щодо вибору методів попередньої обробки в залежності від особливостей текстових даних

Особливості тексту	Методи попередньої обробки	Очікуваний ефект
Розмір від 1 до 3 слів	Токенізація, нормалізація	Слова в тексті приведені до форми основи чи леми, що дає можливість порівнювати їх у векторному представленні зі словами іншого тексту.
Розмір більше 3 слів	Токенізація, нормалізація, видалення стоп-слів	Слова в тексті приведені до форми основи чи леми, що дає можливість порівнювати їх у векторному представленні зі словами іншого тексту. Відсутність стоп-слів дозволяє скоротити розмір векторного представлення тексту.
Наявність абревіатур, скорочень	Розкриття абревіатур, скорочень	Дозволяє отримати однакові фрагменти векторного представлення текстів, коли в одному з них наявна абревіатура або скорочення, а в іншому – повне визначення.
Помилки друку	Методи виправлення помилок (за наявності відповідних функцій в програмних засобах для заданої мовної моделі)	Дозволяє отримати правильні представлення для помилково записаних слів. Помилки значною мірою можуть бути присутні в пошукових запитах користувачів.
Жаргонізми	Заміна жаргонних слів на еквівалентні (методи на основі словників або машинне навчання)	Дозволяє отримати однакові фрагменти векторного представлення текстів для жаргонних та еквівалентних їм слів. Як правило, жаргонізми присутні в пошукових запитах користувачів.
Текст не містить	Стоп-слова не видаляються	У випадку, коли список стоп-слів містить загальновживану лексику, що є частиною

семантично значущих елементів		пошукового запиту, присутність цих слів у вже обробленому тексті дозволяє врахувати їх при порівнянні текстів. Однак існує небезпека формування занадто великого списку релевантних відповідей.

Зазначені методи можуть застосовуватись незалежно один від одного, але послідовність застосування методів має значення для кінцевого результату (наприклад, видалення стоп-слів варто проводити після нормалізації, а нормалізацію після розшифрування абревіатур). Слід зауважити, що тексти, що описують об’єкти інтересу, в освітній сфері, як правило, є текстами високої синтаксичної, граматичної, лексичної та семантичної якості, оскільки проходять багатократну перевірку перед тим, як бути опубліковані на деякому ресурсі. Пошукові запити користувача, в свою чергу, можуть містити значну зашумленість різного характеру.

Висновки по розділу 3

У даному розділі визначено характеристики рекомендаційної системи в сфері освіти, проведено моделювання її структури. Розроблено модель даних рекомендаційної системи в сфері освіти на основі нечіткої логіки, яка лягла в основу розробки методу формування рекомендацій за об’єктами інтересу з використанням гібридного підходу.

Запропонований метод автоматизованого формування рекомендації на основі гібридного підходу дозволяє врахувати особливості характеристик об’єктів інтересу відповідно до предметної галузі формування рекомендації для здобувачів освіти. Застосування попередньої фільтрації об’єктів інтересу з використанням методів обробки текстової інформації дозволяє скоротити розмір вектору потенційно корисних об’єктів. Специфічні для освітнього процесу цільові опитування використовуються з тою ж метою, що дозволяє в

подальшому суттєво зменшити розміри матриць колаборативної фільтрації. Також, рейтингування об’єктів інтересу на основі цільових опитувань дозволяє зменшити проблему «холодного старту», оскільки кількість об’єктів, про які немає рейтингової інформації в системі, є дуже низькою, і навіть для таких об’єктів можна легко отримати попередню оцінку на основі аналогічних оцінок схожих з ними об’єктів.

РОЗДІЛ 4.
ПРАКТИЧНА РЕАЛІЗАЦІЯ ТА ОЦІНКА ЕФЕКТИВНОСТІ МЕТОДИКИ ПОБУДОВИ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ ДЛЯ СФЕРИ ОСВІТИ НА ОСНОВІ ГІБРИДНОГО ПІДХОДУ

4.1 Визначення параметрів моделей та методів рекомендаційної системи для формування рекомендації навчальної дисципліни та літератури

Практична побудова рекомендаційної системи на основі розробленої методики передбачає проведення аналізу особливостей предметної галузі відповідно до кроків, описаних в пункті 3.5, та визначення параметрів моделей та методів, що будуть використані в рекомендаційній системі.

Першою групою об’єктів інтересу в контексті даної задачі виступають дисципліни вільного вибору, що представлені деяким каталогом. Каталог може бути частиною бази даних системи. Мінімально набір інформації, що необхідний студентові для вибору дисципліни і не залежить від зовнішніх факторів (оцінювання, взаємодії тощо) обмежується наступним списком:

- назва дисципліни;
- семестр (семестри), в якому вона вивчається;
- кількість кредитів, що відводяться на дисципліну;
- форма підсумкового оцінювання (залік чи іспит);
- анотація дисципліни.

Для визначення числового рейтингу дисципліни будемо використовувати підхід, при якому рейтинг формується, як комплексна оцінка на основі декількох факторів. В роботі було використано анкету опитування студентів щодо якості освітнього процесу. Опитування представляло собою Google-форму і містило загальні питання щодо організації освітнього процесу, а також блок оцінювання дисциплін. Кожна дисципліна спочатку оцінювалась чисельно за 8 параметрами, а саме:

- кваліфікація викладача;

– відповідність дисципліни силабусу;
– сучасність навчальних матеріалів;
– відповідність методів навчання;
– цікавість навчального матеріалу;
– складність навчального матеріалу;
– прозорість критеріїв оцінювання;
– об'єктивність оцінювання.

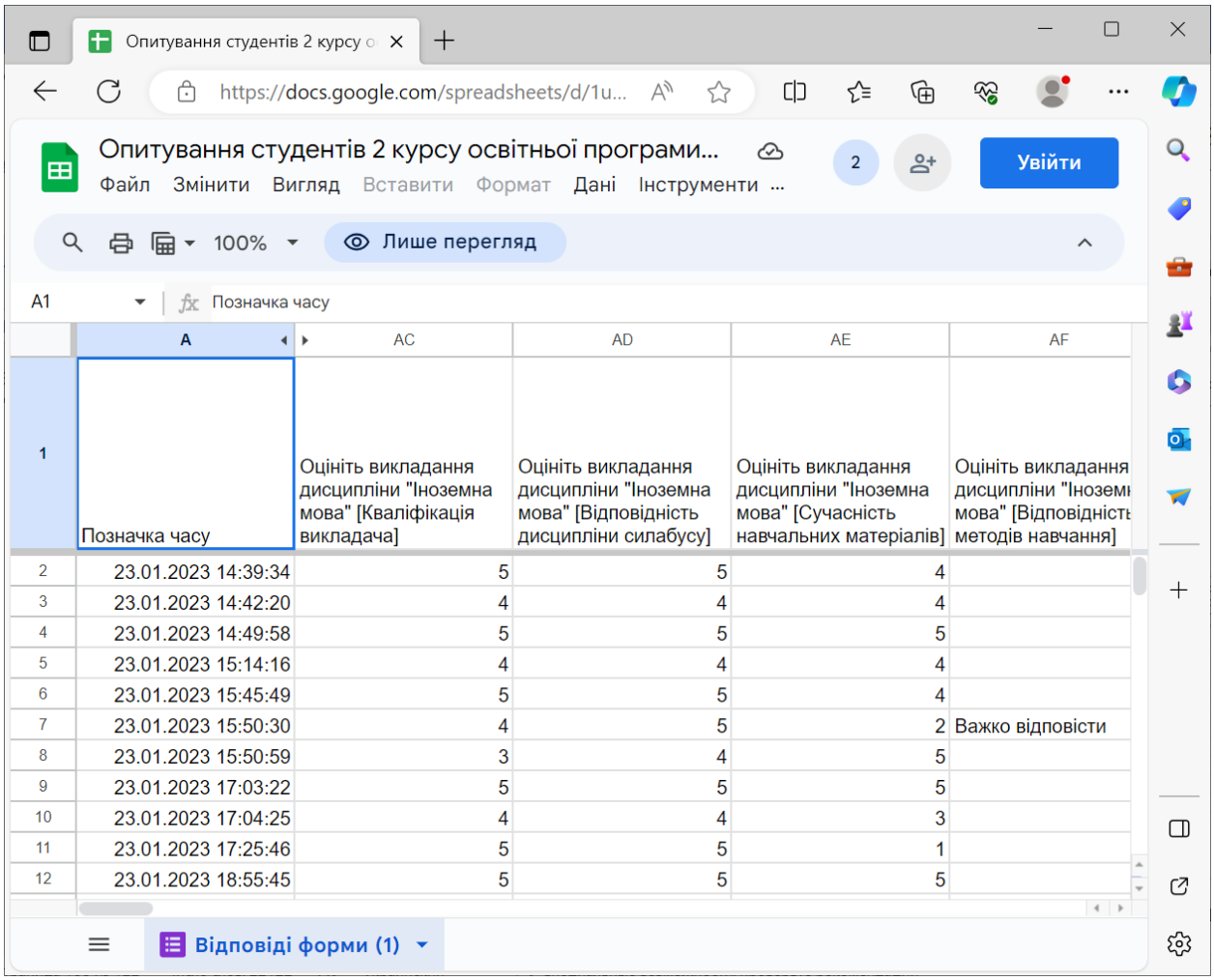
Оцінювання проводилось за шкалою від 1 до 5 з додатковим лінгвістичним значенням «Важко відповісти», яке може інтерпретуватись, як нульове значення, що виключає врахування цього показника в моделі оцінювання.

Додатково до оцінювальних значень студентам пропонувалось зазначити, чи порекомендують вони дисципліну для вивчення наступним групам студентів. Оцінювання по цьому критерію виконувалось в лінгвістичній формі і містило шкалу значень «точно не порекомендую», «скоріше не порекомендую», «скоріше порекомендую», «точно порекомендую». За результатами опитувань автоматично формується таблиця, що має вигляд на рис.4.1. Поля B-AB приховані, оскільки вони містять конфіденційну інформацію, що не має відношення до вирішення задачі побудови рекомендації щодо вибору дисципліни.

Будемо обчислювати частковий зовнішній рейтинг $R1_{jk}$ дисципліни з номером j, наданий k-м студентом, як величину (4.1), що визначається адитивною моделлю і враховує значення показників x_1 «кваліфікація викладача», x_2 «відповідність дисципліни силабусу», x_3 «сучасність навчальних матеріалів», x_4 «відповідність методів навчання», x_5 «цікавість навчального матеріалу», x_6 «складність навчального матеріалу», x_7 «прозорість критеріїв оцінювання», x_8 «об'єктивність оцінювання»:

$$R1_{jk} = \sum_{i=1}^{8} \alpha_i \, (x_i)_{jk} \tag{4.1}$$

де $(x_i)_{jk}$ – оцінка, надана k-м студентом для дисципліни з номером j за показником з номером i; α_i– коефіцієнт важливості показника з номером i в загальній оцінці. Вважливість показника визначається однаковою для всіх дисциплін та всіх студентів, що оцінювали дисципліни. Для забезпечення нормованості величини $R1_{jk}$ коефіцієнти α_i в сумі повинні становити 1.



	A	AC	AD	AE	AF
1	Позначка часу	Оцініть викладання дисципліни "Іноземна мова" [Кваліфікація викладача]	Оцініть викладання дисципліни "Іноземна мова" [Відповідність дисципліни силабусу]	Оцініть викладання дисципліни "Іноземна мова" [Сучасність навчальних матеріалів]	Оцініть викладання дисципліни "Іноземн мова" [Відповідність методів навчання]
2	23.01.2023 14:39:34	5	5	4	
3	23.01.2023 14:42:20	4	4	4	
4	23.01.2023 14:49:58	5	5	5	
5	23.01.2023 15:14:16	4	4	4	
6	23.01.2023 15:45:49	5	5	4	
7	23.01.2023 15:50:30	4	5	2	Важко відповісти
8	23.01.2023 15:50:59	3	4	5	
9	23.01.2023 17:03:22	5	5	5	
10	23.01.2023 17:04:25	4	4	3	
11	23.01.2023 17:25:46	5	5	1	
12	23.01.2023 18:55:45	5	5	5	

Рис. 4.1. Результати семестрового цільового опитування студентів

Представимо коефіцієнт (4.1) нечіткою змінною $\widetilde{R1_{jk}}$ з терм -множиною:

"оцінка дисципліни"=<"дуже низька оцінка", "низька оцінка", "середня оцінка", "висока оцінка", "дуже висока оцінка">, (4.2)

і трикутними функціями належності (4.3)-(4.7):

$$\mu_{\text{дуже низька оцінка}}(R1_{jk}) = \begin{cases} 1 - R1_{jk}, \text{ при } 0 < R1_{jk} \leq 1 \\ 0, \text{ інакше} \end{cases}; \quad (4.3)$$

$$\mu_{\text{низька оцінка}}(R1_{jk}) = \begin{cases} R1_{jk} - 1, \text{при } 1 < R1_{jk} \leq 2 \\ 3 - R1_{jk}, \text{ при } 2 < R1_{jk} \leq 3; \\ 0, \text{ інакше} \end{cases} \quad (4.4)$$

$$\mu_{\text{середня оцінка}}(R1_{jk}) = \begin{cases} R1_{jk} - 2, \text{при } 2 < R1_{jk} \leq 3 \\ 4 - R1_{jk}, \text{ при } 3 < R1_{jk} \leq 4; \\ 0, \text{ інакше} \end{cases} \quad (4.5)$$

$$\mu_{\text{висока оцінка}}(R1_{jk}) = \begin{cases} R1_{jk} - 3, \text{при } 3 < R1_{jk} \leq 4 \\ 5 - R1_{jk}, \text{ при } 4 < R1_{jk} \leq 5; \\ 0, \text{ інакше} \end{cases} \quad (4.6)$$

$$\mu_{\text{дуже висока оцінка}}(R1_{jk}) = \begin{cases} R1_{jk} - 4, \text{ при } 4 < R1_{jk} \leq 5 \\ 0, \text{ інакше} \end{cases}. \quad (4.7)$$

Загальна нечітка оцінка $\widetilde{R1_j}$ по всім студентам може бути обчислена з використанням будь якої операції з нечіткими множинами (перетин, добуток, об’єднання, сума).

Нечітку рекомендацію, що студент надає наступним студентам позначимо змінною $\widetilde{R2_{jk}}$ з терм -множиною (4.8):

"рекомендація студента"=<"точно не порекомендую", "скоріше не порекомендую", "скоріше порекомендую", "точно порекомендую">. (4.8)

Кожній лінгвістичній рекомендації студента $\widetilde{R2_{jk}}$ поставимо у відповідність точні числа $R2_{jk} \in \{0, 1, 2, 3\}$. Загальна рекомендація дисципліни $R2_j$ по всім студентам буде визначатиметься як середнє значення (4.9):

$$R2_j = \sum_k^K R2_{jk}, \tag{4.9}$$

де K – загальна кількість студентів, що приймали участь в опитуванні.

Переведення чіткого числа $R2_j$ в лінгвістичне представлення $\widetilde{R2}_j$ проведемо з використанням терм-множини (4.8) з функціями належності (4.10)-(4.13) і універсумом [1; 4]:

$$\mu_{\text{точно не порекомендую}}(R2_j) = \begin{cases} 1 - R2_j, \text{ при } 0 < R2_j \leq 1 \\ 0, \text{ інакше} \end{cases}; \tag{4.10}$$

$$\mu_{\text{скоріше не порекомендую}}(R2_j) = \begin{cases} R2_j - 1, \text{при } 1 < R2_j \leq 2 \\ 3 - R2_j, \text{ при } 2 < R2_j \leq 3; \\ 0, \text{ інакше} \end{cases} \tag{4.11}$$

$$\mu_{\text{скоріше порекомендую}}(R2_j) = \begin{cases} R2_j - 2, \text{при } 2 < R2_j \leq 3 \\ 4 - R2_j, \text{ при } 3 < R2_j \leq 4; \\ 0, \text{ інакше} \end{cases} \tag{4.12}$$

$$\mu_{\text{точно порекомендую}}(R2_j) = \begin{cases} R2_j - 3, \text{ при } 3 < R2_j \leq 4 \\ 0, \text{ інакше} \end{cases}. \tag{4.13}$$

Загальний рейтинг дисципліни на основі рекомендацій та оцінок студентів представимо лінгвістичною змінною R_j з терм-множиною (4.14)

"загальний рейтинг"=<"дуже низький", "низький", "середній", "вище середнього", "високий">, (4.14)

і функціями належності (4.15)-(4.19) з універсумом [0; 1], поділеним на 5 діапазонів з кроком 0,25:

$$\mu_{\text{дуже низький}}(x) = exp\left[-\left(\frac{x}{\delta}\right)^2\right], \tag{4.15}$$

$$\mu_{\text{низький}}(x) = exp\left[-\left(\frac{x-0.25}{\delta}\right)^2\right], \tag{4.16}$$

$$\mu_{\text{середній}}(x) = exp\left[-\left(\frac{x-0.5}{\delta}\right)^2\right], \quad (4.17)$$

$$\mu_{\text{вище середнього}}(x) = exp\left[-\left(\frac{x-0.75}{\delta}\right)^2\right], \quad (4.18)$$

$$\mu_{\text{високий}}(x) = exp\left[-\left(\frac{x-1}{\delta}\right)^2\right], \quad (4.19)$$

де δ – крутизна функції належності, для адекватного представлення кривих функцій належності крутизну бажано визначати менше 1.

Для обчислення лінгвістичного значення загального рейтингу R_j на основі величин $R1_j$ та $R2_j$ можна застосовувати продукційні правила, описані в розділах 2 та 3, але в цій задачі використаємо обчислення нечіткого числового значення через операцію суми двох нечітких множин (4.20) і по ньому будемо визначати на основі функції належності відповідний терм, як максимум серед значень функцій (4.15)-(4.19):

$$\mu_{R_j}(A_j, B_j) = \mu_{R1_j+R2_j}(A_j, B_j) = \mu_{R1_j}(A_j) + \mu_{R2_j}(B_j) - \mu_{R1_j}(A_j)\mu_{R2_j}(B_j), \quad (4.20)$$

де A_j – матриця всіх оцінок студентів, що стосуються якості викладання j-ї дисципліни, B_j – вектор рекомендацій студентів щодо майбутнього вивчення викладання j-ї дисципліни.

Оскільки зазначені оцінки та рекомендації зберігаються в зовнішньому джерелі (таблиці Google), то в структурі системи необхідний модуль для обробки даних цієї таблиці та модуль інтерфейсу адміністратора системи для встановлення налаштувань параметрів нечітких змінних.

Щодо текстової інформації про дисципліну, то умовно її можна розділити на 2 блоки: назва та анотація. Обидва тексти є текстами доволі високої якості, оскільки:

– доля синтаксичних, лексичних, семантичних чи інших видів шумів в них дуже низька, завдяки тому, що вони декілька разів перевіряються перед

завантаженням на певний веб-ресурс;

– абревіатури та скорочення як правило наводяться вже розшифрованими; виключення можуть складати специфічні скорочення, що зустрічаються в IT-сфері, але, як правило, вони є сталими абревіатурами і студенти в пошукових запитах використовують саме скорочення, а не повні назви технологій, фреймворків та інших об’єктів.

Таким чином, для зазначених текстових блоків доцільно застосувати легку попередню обробку, що включає операції токенізації, нормалізації, видалення стоп-слів та векторизації.

Інформація про студентів, що використовують рекомендаційну систему, може бути отримана з баз даних деканату чи дистанційної системи навчання. При цьому, безпосередньо для роботи рекомендаційної системи необхідна лише процедура аутентифікації студента, оскільки авторизація (шляхом виділення дійсним студентам закладу освіти логінів та паролів) відбувається на інших рівнях організації освітнього процесу.

Безпосередня зацікавленість студента деякою дисципліною ґрунтується на сфері його інтересів і може бути визначена на основі його пошукових запитів до системи. В процесі роботи системи зберігається вся історія пошукових запитів студента, щоб осягнути максимальну кількість його інтересів. Зважаючи на те, що вибіркові дисципліни студент шукає починаючи не раніше ніж з 2 курсу бакалаврату і виконує цю операцію один чи два рази на рік в залежності від особливостей організації освітнього процесу в закладі, то, наприклад, в середньому кожному студенту бакалаврату може відповідати від 6-10 запитів за період навчання в бакалавраті. Оскільки користувацькі пошукові запити відносяться до текстів низької якості, то вони потребують посиленої попередньої обробки, яка включає в себе наступні операції:

– очищення від знаків пунктуації та спецсимволів;

– токенізація;

– стемінг (для визначення основ слів, що є жаргонізмами чи сленгом, оскільки ймовірність присутності в словниках лем подібних токенів дуже

низька);

– заміна жаргонізмів та сленгових виразів (в роботі використано словник жаргонізмів та сленгових виразів для IT-сфери, в рамках інших спеціальностей експертом повинен бути підготовлений інший словник);

– лематизація;

– видалення стоп-слів;

– векторизація.

Подібність текстів, що визначають користувацьку сферу інтересів до вмісту опису навчальних дисциплін будемо визначати через поняття релевантності і застосовувати підхід на основі вирішення задачі тематичного пошуку. Для цього кожній дисципліні поставимо у відповідність індекси, що будуть відображати зміст тестового опису. В якості індексів оберемо ключові слова документу, які обчислимо на основі статистичної міри TF-IDF. TF-IDF передбачає оцінку важливості слова в контексті документа, що є частиною колекції документів або корпусу. Оскільки одній дисципліні в нашому випадку відповідає лише один опис, то будемо вважати документом окреме речення опису, тоді весь опис – це текстовий корпус, що відповідає певній дисципліні. Таким чином, представимо опис дисципліни як множину D (4.21):

$$D = \{d_1, \ \dots, \ d_{|D|}\}, \tag{4.21}$$

де $d_j \in D$ – окреме речення опису.

Вагу $tfidf$ деякого терму $t_i \in d_j$ будемо обраховувати за виразом (4.22):

$$tfidf(t_i, d_j, D) = tf(t_i, d_j) \cdot idf(t_i, D), \tag{4.22}$$

де $tf(t_i, d_j)$ – частота вживання слова в документі, $tf(t_i, d_j) = n_{t_i}/|d_j|$, n_{t_i} – кількість входжень слова в речення, $|d_j|$ - загальна кількість слів в реченні; $idf(t_i, D)$ – обернена частота вживання слова в усіх документах колекції, $idf(t_i, D) = \frac{|D|}{|\{d_i \in D | t_j \in d_i\}|}$, $|\{d_i \in D | t_j \in d_i\}|$ - кількість речень в описі D, в яких

зустрічається слово t_j. До пошукового індексу $D`$ включаються всі слова опису $D$, для яких $tfidf(t_i, d_j, D) > \lambda$, де $\lambda$ – поріг відбору. Ступінь релевантності пошукового запиту $S$ користувача до опису дисципліни $D$ визначається на основі знайдених співпадінь слів пошукового запиту $s_k \in S$ з елементами пошукового індексу $D`$ (4.23):

$$\rho(S, D`) = \frac{|s_k \in D`|}{|D`|}. \quad (4.23)$$

Значення (4.23) обчислюється для кожного опису дисципліни, отримані значення впорядковуються за зростанням. Дисципліни, чия релевантність описів перевищує деякий поріг P, включаються в підмножину для подальшого відбору для рекомендації.

Для підвищення впливу користувацької взаємодії з об’єктами інтересу будемо додатково використовувати семестрові оцінки, які отримали студенти, що вже вибрали та пройшли відповідний курс дисципліни. Ця інформація може бути отримана з баз даних деканату. При наявності автоматизованої системи обліку успішності студентів додаткової особистої інформації, окрім логіну та паролю для аутентифікації студента, може не бути затребувано. У випадку відсутності засобів автоматизованого обліку студентів є необхідність зберігати в базі інформацію про ПІП студента для можливості внесення його оцінок по дисциплінам. Вагу дисципліни в контексті семестрових оцінок будемо обчислювати як нормоване число в діапазоні [0, 1] за формулою (4.24):

$$RM_j = \frac{\sum_{i=1}^{M} mark_{ij}}{M \cdot 100}, \quad (4.24)$$

де $mark_{ij}$ – оцінки користувачів з j-ї дисципліни, M– загальна кількість оцінок з дисципліни, число 100 визначає максимальне значення шкали оцінювання семестрової оцінки. Будемо вважати, що якщо нормована вага дисципліни в контексті семестрових оцінок знаходиться в діапазоні [0; 0,65], то це означає, що студенти погано засвоїли дисципліну (можливо, із-за складності матеріалу,

можливо із-за відсутності інтересу тощо їх середній бал не перевищує поріг слабкої трійки в національній шкалі оцінювання) і рекомендація такої дисципліни повинна бути скоригована в сторону зменшення ваги. Якщо вага дисципліни в контексті семестрових оцінок знаходиться в діапазоні [0,9; 1], то це може означати, що середня оцінка з дисципліни становить відмінно за національною шкалою, а це є сигналом до необ'єктивного оцінювання або занадто легкого матеріалу. Такі дисципліни в рекомендацію попадати не будуть. Ваги в діапазоні (0,65; 0,9) будуть враховані для підвищення рейтингу дисципліни.

Задача визначення подібності навчально-методичної літератури до пошукового запиту користувача по своїм характеристикам подібна до задачі визначення подібності пошукового запиту користувача до опису дисципліни, тому вирішується схожим чином: в якості текстового корпусу виступає літературне джерело, однак, в якості елементів (документів), на основі яких ведеться пошук ключових слів можуть виступати не окремі речення, а, наприклад, абзаци чи розділи (глави) джерела. Крім того, в силу того, що цей контент не підлягає ніяким змінам в процесі роботи рекомендаційної системи, списки ключових слів, що представляють конкретне джерело, можуть бути побудовані окремо поза межами системи (в тому числі, не автоматичними способами, а за участі експертів) та використовуватись в ній у вигляді вже готової бази даних.

Для врахування дій користувача будемо використовувати колаборативний метод, який ґрунтується на порівнянні між собою об'єктів інтересу користувачів. Алгоритм побудови рекомендації включає наступні кроки:

1. Для кожного літературного джерела j обчислюється, наскільки воно схоже з джерелом i, для якого виконується передбачення його рейтингової оцінки (рангу).

2. Обирається множина літературних джерел, максимально близьких до i-го.

3. Оцінка передбачається на основі оцінок, що виставлені користувачем u для множини, визначеної на 2 кроці цього алгоритму.

Для зменшення обчислювальної складності кроку 1 всю множину літературних джерел варто розбити на окремі кластери та обчислювати відстані тільки для елементів, що знаходяться в одному з ним кластері. При додаванні нового джерела в каталог для нього повинна бути визначена належність до одного з вже наявних кластерів. Загальна задача кластеризації літературних джерел можу бути визначена наступним чином. Є вхідна множина літературних джерел (4.25)

$$L = \{l_1, \ \dots, \ l_{|L|}\}. \tag{4.25}$$

Передбачається, що можна автоматично розбити L на кластери і вони будуть відповідати певним категоріям C = $\{c_1, \ \dots, \ c_{|C|}\}$. Задача зводиться до пошуку такого C, яке було б оптимальним у відповідності до деякого критерію якості розбиття. До кластерів зазвичай висуваються наступні вимоги: схожість між об’єктами з одного кластеру повинна бути якомога більшою; схожість між об’єктами різних кластерів повинна бути якомога меншою.

Для визначення схожості між об’єктами будемо використовувати косинусну відстань між векторними представленнями текстів літературних джерел :

$$cos_sim(l_i, l_j) = \frac{\sum_{k=1}^{n} l_{i_k} l_{j_k}}{\sqrt{\sum_{k=1}^{n} l_{i_k}^2}\sqrt{\sum_{k=1}^{n} l_{j_k}^2}}, \tag{4.26}$$

де l_i, l_j – відповідні мішки слів джерел з номерами i та j, k – номер слова в мішку слів, n – розмір мішка слів.

Кластеризацію літературних джерел будемо виконувати з використанням плаского нечіткого алгоритму c-середніх (c-means), який є модифікацією методу k-means. Послідовність кроків для виконання клатеризації наведена нижче:

1. Вибираємо k початкових центроїдів кластерів.

2. Кожен документ відносимо до m кластерів, чиї центроїди є найближчими.

3. Обчислюємо нові центроїди кластерів.

4. Повторюємо кроки 2-3 доки не буде досягнуто умову закінчення.

Особливістю цього методу є те, що він дозволяє обробити ситуацію належності елементу одночасно до декількох кластерів. Це є важливим фактором, оскільки одне й те саме літературне джерело може бути використано при вивченні різних дисциплін. Таким чином, при виконанні алгоритму побудови рекомендації в кроці 1 порівнюються всі джерела, що лежать в тих самих кластерах, що і джерело, для якого будується передбачення.

Оскільки літературні джерела є текстами доволі високої семантичної, лексичної та синтаксичної якості, то попередня обробка текстів літературних джерел зводиться до набору наступних операцій:

– очищення від знаків пунктуації, спецсимволів, цифр;

– приведення до нижнього регістру;

– токенізація;

– нормалізація (можна використовувати як стемінг для збільшення швидкості обробки, так і лематизацію для підвищення якості токенів, але це збільшить складність попередньої обробки та час її виконання);

– видалення стоп-слів (повинно виконуватись обов’язково, при чому, стандартні словники стоп-слів за необхідності можуть бути розширені словами, що не несуть семантичного навантаження в рамках окремої галузі освіти, спеціальності чи освітньої програми)

– векторизація.

4.2 Розробка структур баз даних рекомендаційної системи для формування рекомендації навчальної дисципліни та літератури

Для представлення даних системи було розроблено відповідні структури баз даних, представлені схемою на рис.4.2.

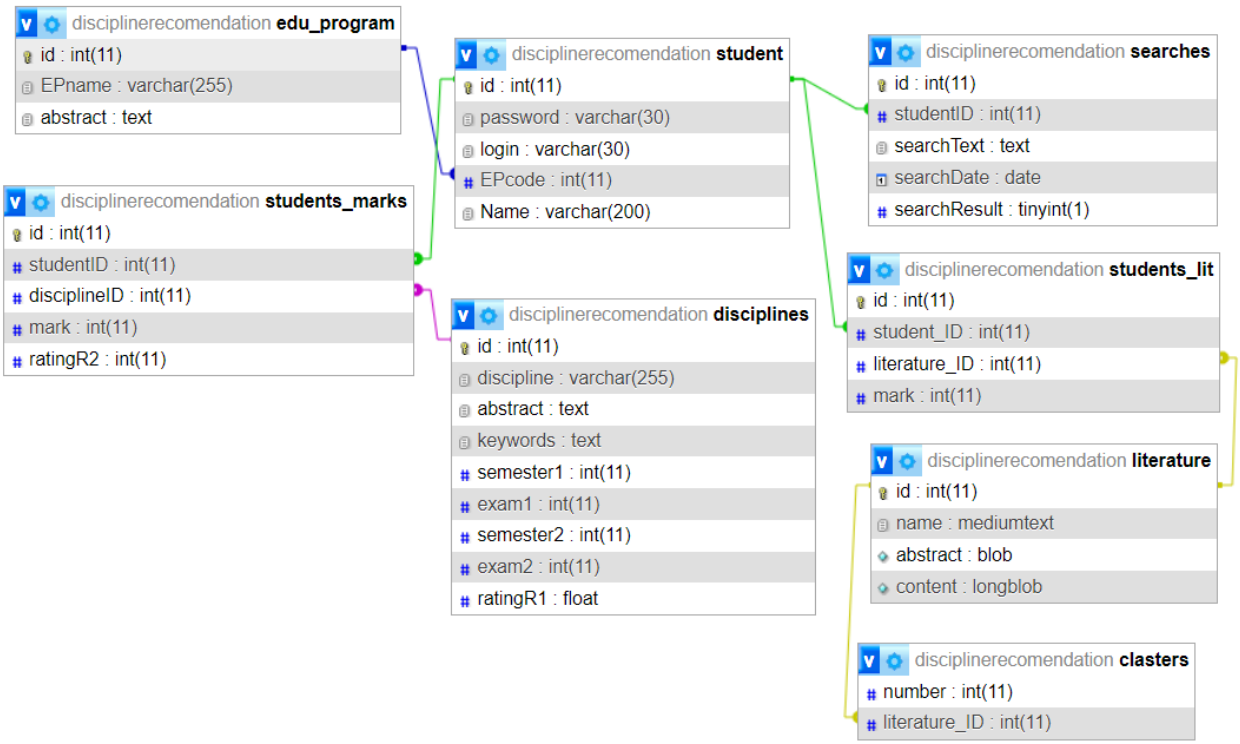


Рис.4.2. Структура бази даних для задачі формування рекомендацій щодо вибору дисципліни та підбору літератури

Таблиця edu_program призначена для зберігання інформації про освітню програму, на якій навчається студент, та містить наступні поля:

– id - первинний ключ, код освітньої програми, поле має цілочисельний тип, унікальний ідентифікатор освітньої програми, який автоматично збільшується (AUTO_INCREMENT);

– EPname - рядкове поле, що містить назву освітньої програми, поле є обов'язковим для заповнення;

– Abstract - текстове поле, яке містить короткий опис освітньої програми, поле є обов'язковим для заповнення;

Таблиця student призначена для зберігання інформації про студента, та містить наступні поля:

– id – первинний ключ, поле має цілочисельний тип, унікальний ідентифікатор студента, який автоматично збільшується (AUTO_INCREMENT);

– Name - рядкове поле, що містить ім'я та прізвище студента (може бути відсутнім або закодованим);

– EPcode - поле, яке представляє код освітньої програми, воно є зовнішнім ключем, який посилається на поле id в таблиці OP, поле має цілочисельний тип;

– password -поле текстового типу, призначено для зберігання паролю для входу студента в систему, оскільки система не містить конфіденційної інформації, то ніяких заходів для захисту інформації в системі не передбачено;

– login – поле текстового типу, логін користувача системи; і логін і пароль, як було зазначено раніше, не формуються всередині системи, а передаються до неї ззовні, таким чином користувач може тільки проходити аутентифікацію в системі.

Таблиця searches призначена для зберігання інформації про пошукові запити студента в контексті вибору дисципліни і містить наступні поля:

– id - первинний ключ таблиці, цілочисельний тип, унікальний ідентифікатор пошукового запиту, який автоматично збільшується (AUTO_INCREMENT);

– studentID - код студента, цілочисельний тип, обов'язкове поле, є зовнішнім ключем, який посилається на поле id в таблиці student;

– searchText - текстове поле великого розміру, яке містить вміст пошукового запиту;

– searchDate – поле типу дата, використовується для збереження часу звернення користувача до системи, створено для забезпечення можливості відбору більш актуальних запитів користувача, оскільки з часом його сфера

інтересів може змінитись, також цей параметр дозволяє визначати активних та пасивних користувачів протягом певного проміжку часу;

– searchResult - ціле число, яке представляє оцінку релевантності результату пошуку, це поле є необов'язковим і може залишатися порожнім (NULL), якщо студент не захотів оцінити результати роботи рекомендаційної системи, поле є числовою інтерпретацією нечіткої лінгвістичної змінної.

Таблиця disciplines призначена для зберігання даних про дисципліну, яка пропонується для вивчення студентам, та містить поля:

– id - код дисципліни, первинний ключ таблиці, цілочисельний тип, унікальний ідентифікатор дисципліни, який автоматично збільшується (AUTO_INCREMENT);

– discipline - рядкове поле, що містить назву дисципліни, обов'язкове для заповнення;

– abstract - текстове поле, що містить короткий опис дисципліни, обов'язкове для заповнення;

– keywords – текстове поле, що містить ключові слова дисципліни, може формуватись засобами системи або мати зовнішнє джерело для заповнення (наприклад, дані з робочої програми чи силабусу, експертні думки викладача тощо);

– semester1 – номер семестру, в якому вивчається дисципліна, обов'язкове для заповнення, оскільки дисципліна повинна вивчатись хоча б один семестр, має цілочисельний тип;

– exam1 – форма оцінювання дисципліни в першому семестрі, має цілочисельний тип, приймає одне із двох значень: 1 - залік, 2 - екзамен, обов'язкове для заповнення;

– semester2 – цілочисельне поле, що представляє номер другого семестру, в якому вивчається дисципліна, поле є необов'язковим, оскільки другий семестр при вивченні дисципліни може бути відсутній (в системі враховуються лише дисципліни, що вивчаються не більше 2 семестрів,

оскільки, як правило, дисципліни, що вивчаються більшу кількість семестрів, є обов’язковими та не є об’єктом інтересу даної системи);

– exam2 - форма оцінювання дисципліни в другому семестрі вивчення, має цілочисельний тип, приймає одне із двох значень: 1 - залік, 2 - екзамен, необов'язкове для заповнення;

– ratingR1 – рейтинг дисципліни $R1$, обчислений відповідно до наведеної в цьому розділі моделі, числове поле.

Таблиця students_marks призначена для зберігання оцінок студентів з дисципліни у випадку, якщо студент її вивчав. Таблиця містить наступні поля:

– id - первинний ключ таблиці, цілочисельний тип, унікальний ідентифікатор оцінки, який автоматично збільшується (AUTO_INCREMENT);

– studentID - код студента, цілочисельне поле, зовнішній ключ, який посилається на поле id в таблиці student;

– disciplineID - код дисципліни, цілочисельне поле, зовнішній ключ, який посилається на поле id в таблиці disciplines;

– mark - оцінка, яку студент отримав з дисципліни, якщо вивчав її, цілочисельне поле, має значення в діапазоні від 0 до 100;

– ratingR2 – значення, що використовується для обчислення рейтингу дисципліни $R2$, числова інтерпретація рекомендації студента щодо вивчення даної дисципліни, числове поле.

Слід зазначити, що для виставлення рейтингу студент не обов’язково повинен вивчати дисципліну – думка щодо рейтингу може сформуватись на основі аналізу наявної у студента інформації про дисципліну (відгуки інших студентів, враження про викладача дисципліни, аналіз анотації дисципліни та інші фактори).

Таблиця literature призначена для зберігання даних про літературні джерела, які пропонуються студентам, та містить поля:

– id - код літературного джерела, первинний ключ таблиці, цілочисельний тип, унікальний ідентифікатор джерела, який автоматично збільшується (AUTO_INCREMENT);

– name - рядкове поле, що містить назву джерела, обов'язкове для заповнення, в загальному випадку зазначена назва може біти як повним бібліографічним описом, так і скороченим варіантом, що включає лише список авторів та безпосередньо назву;

– abstract - текстове поле, що містить короткий опис літературного джерела; це може бути анотація джерела зі списком чи без списку ключових слів, вступ чи частковий фрагмент, що націлений на ознайомлення читача з джерелом, загальних вимог до вмісту цього поля з точки зору семантичного наповнення немає, тому воно використовується виключно в інтерфейсних цілях і не приймає участь у формуванні рекомендації, однак, за необхідності поле може бути об’єднано з основним контентом та використовуватись для побудови рекомендації;

– content – текстове поле, що містить основний текст джерела, дані в це поле повинні імпортуватись з вхідного файлу, тому при наповненні бази треба враховувати доступність текстового представлення джерела (наприклад, файли з захищеним вмістом неможливо буде обробляти в системі);

Таблиця clasters призначена для зберігання даних про результати кластеризації літературних джерел, та містить поля:

– number – числове поле, що визначає номер кластеру, до якого відноситься літературне джерело (одне й те саме джерело може належати до різних кластерів);

– literature_ID – числове поле, зовнішній ключ з таблиці literature, визначає код літературного джерела, якому співставлено кластер з номером number.

Визначені в цьому пункті таблиці дозволяють зберегти дані, що є внутрішніми по відношенню до рекомендаційної системи. Частина даних повинна вводитись оператором системи, частина розраховується завдяки алгоритмам, які є частиною методики, частина полів формується на основі зовнішніх даних.

4.3 Моделювання архітектури та реалізація рекомендаційної системи для формування рекомендації навчальної дисципліни та літератури

Архітектуру рекомендаційної системи для формування рекомендації навчальної дисципліни та літератури (рис.4.3) побудовано на основі структурної моделі, визначеної в розділі 3 та з урахуванням особливостей формування даних для рекомендації дисципліни і літературних джерел.

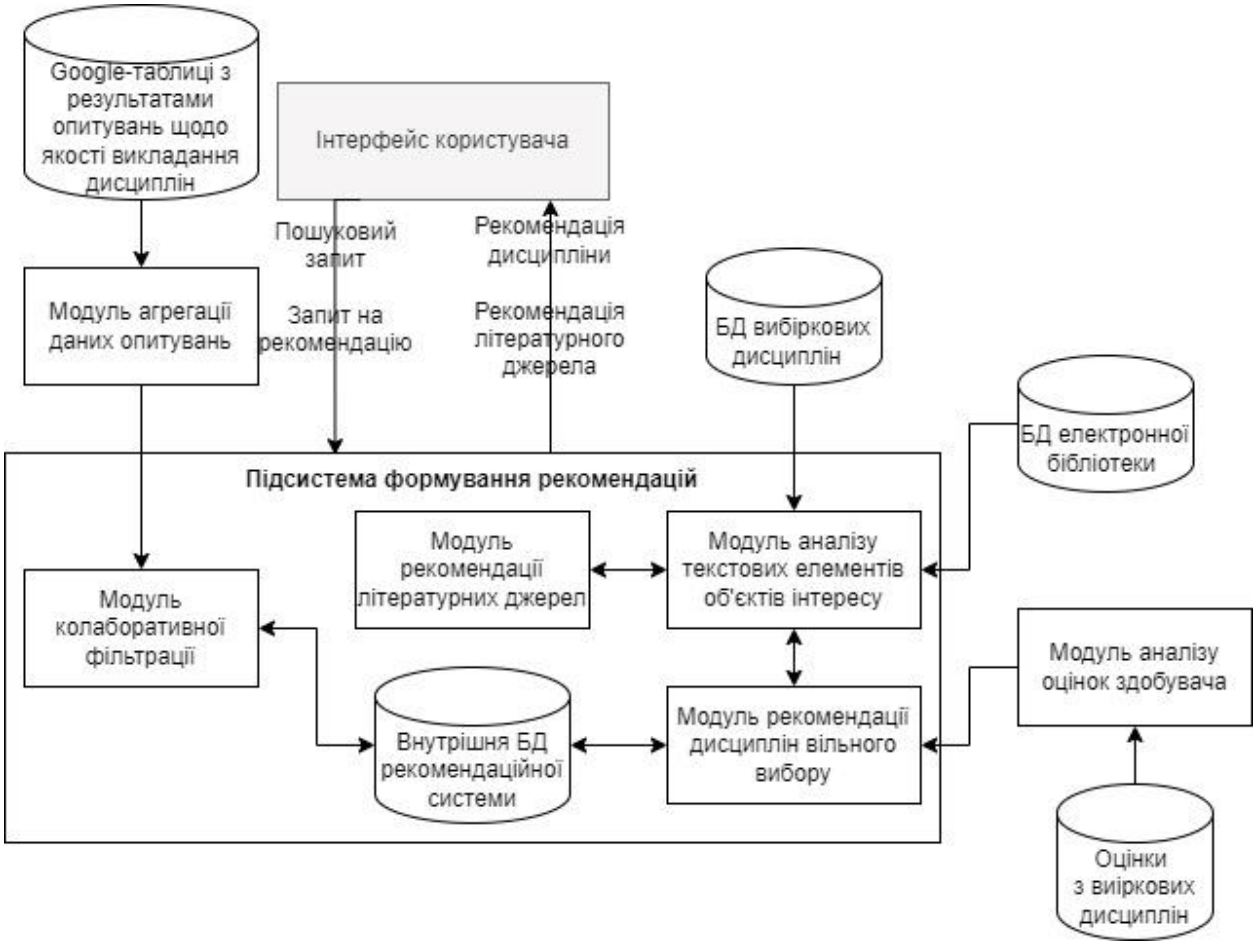


Рис.4.3. Архітектура рекомендаційної системи для формування рекомендації навчальної дисципліни та літератури

На рис. 4.4-4.9 наведено приклади екранних форм системи в режимі користувача. Головна сторінка системи (рис. 4.4) містить навігаційне меню для доступу до каталогу дисциплін та каталогу літератури, кнопку авторизації для входу в акаунт та кнопку/перемикач «Рекомендація», яка дозволяє

отримати рекомендацію по заданому об’єкту інтересу. Як було описано раніше, для формування рекомендації використовуються заздалегідь завантажені результати цільових опитувань студентів. Таким чином, з точки зору користувача система виглядає «чорним ящиком» і весь механізм формування рекомендації від кінцевого користувача прихований. Новини є додатковим елементом і демонструють нові надходження до каталогів.

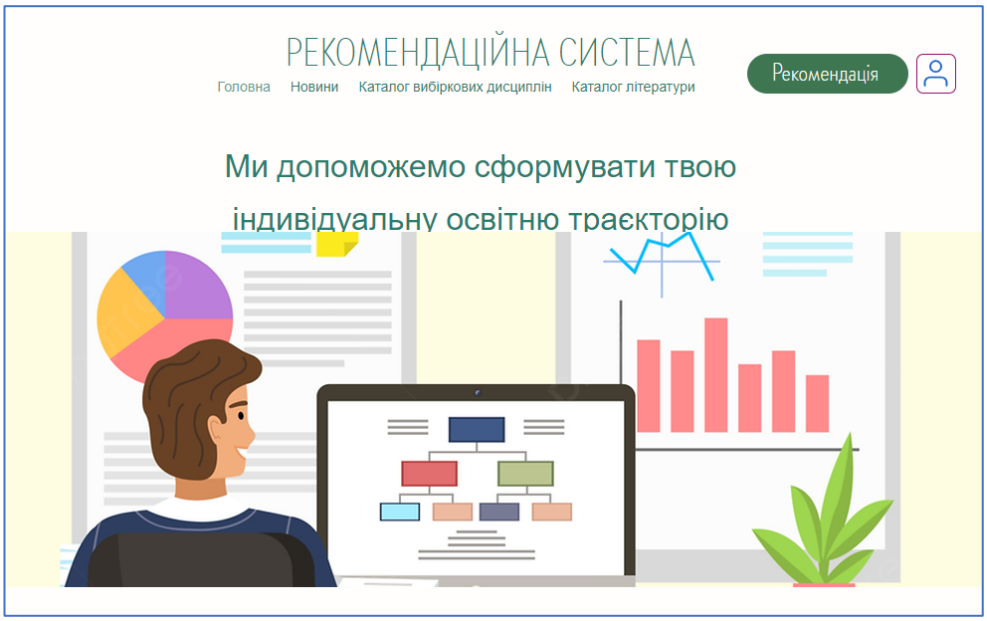


Рис.4.4. Головна сторінка сайту рекомендаційної системи

Рис.4.5. Фрагмент сторінки новин з рекламою нових об’єктів інтересу

Рис. 4.6. Фрагмент сторінки каталогу вибіркових дисциплін

Рис. 4.7. Фрагмент сторінки каталогу вибіркових дисциплін після надання автоматичної рекомендації

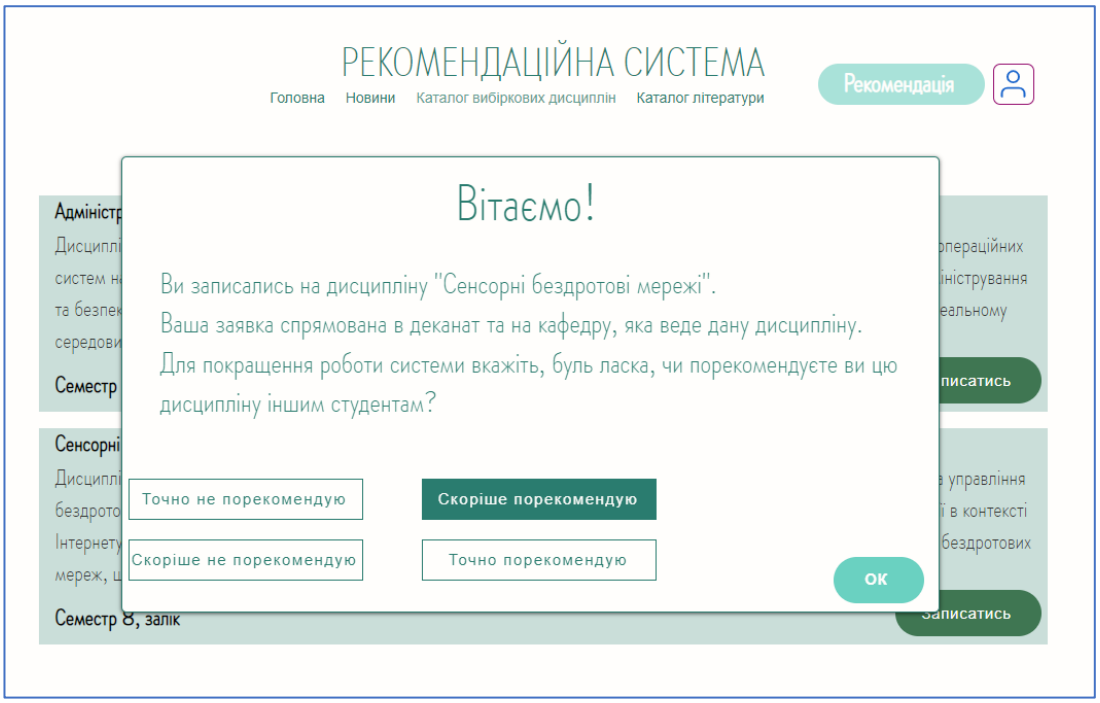


Рис. 4.8. Фрагмент сторінки каталогу вибіркових дисциплін після надання автоматичної рекомендації з пропозицією надати рекомендацію

4.4 Оцінка ефективності розробленої методики

Оцінка якості рекомендаційних систем в першу чергу залежить від цілей їх створення та впровадження. Переважна більшість рекомендаційних систем призначена для вирішення задач електронної комерції – надати рекомендацію, щоб в подальшому продати якийсь продукт чи послугу. Освітня сфера має інший фокус, тому розповсюджені підходи до оцінювання на основі очікуваного прибутку, розширення воронки конверсії та інші подібні недоцільно застосовувати для оцінки рекомендаційної системи в сфері освіти.

Якщо розглядати оцінку якості рекомендаційної системи в контексті підтримки прийняття рішень, то користувач обирає чи не обирає рекомендований йому елемент і може оцінити своє рішення за наступними параметрами [164, 165, 166]:

– істинно-позитивний результат (True Positive, TP) – об’єкт інтересу був рекомендований системою та обраний користувачем;

– хибно-позитивний результат (False Positive, FP) – об’єкт інтересу був рекомендований системою, але користувач його не обрав;

– хибно-негативний результат (False Negative, FN) – об’єкт інтересу не був рекомендований системою, але користувач його все одно обрав;

– істинно- негативний результат (True Negative, TN) – об’єкт інтересу не був рекомендований системою, і користувач теж його не вибрав.

Зазначені параметри можна використати в метриках точності та відгуку (4.26), (4.27):

$$\text{Точність } = \frac{\text{TP}}{\text{TP+FP}}; \tag{4.26}$$

$$\text{Відгук} = \frac{\text{TP}}{\text{TP+FN}}. \tag{4.27}$$

Точність демонструє частку тих об’єктів інтересу, що визначені правильно, серед усіх рекомендованих. Оскільки при побудові рекомендації для вибору дисципліни, список рекомендованих дисциплін не дуже часто змінюється, а протягом навчального року величина $\text{TP} + \text{FP}$ взагалі є константою, то використання показнику точності для оцінки якості роботи рекомендаційної системи не дає корисної інформації для подальшої організації навчального процесу. Крім того, система повинна надавати заздалегідь більше рекомендацій, чим насправді може вибрати користувач. Фактично, на кожні ТОП-3 рекомендації користувач повинен вибрати лише одну, яка визначає його подальшу освітню траєкторію – на відміну від інтернет-магазинів, медійних майданчиків та інших ресурсів, де у користувача є можливість обирати одні й ті ж об’єкти інтересу декілька разів, освітній процес при виборі дисципліни передбачає тільки одиничний вибір.

Для задачі побудови рекомендації при виборі навчальної дисципліни більш важливим є показник відгуку, оскільки він показує частку правильно рекомендованих системою дисциплін серед усіх тих, що обрав користувач. Значення відгуку близьке до 1 збільшує ймовірність того, що ще до

безпосереднього вибору користувачами навчальних дисциплін можна спрогнозувати їх список, що дає можливість заздалегідь спланувати навчальне навантаження викладачів та студентів, скласти розклад та виконати інші організаційні заходи з підготовки навчального процесу.

Було проведено моделювання роботи рекомендаційної системи по формуванню рекомендації вибіркової дисципліни та літературних джерел на реальних користувачах в рамках організації освітнього процесу в Державному університеті інформаційно-комунікаційних технологій. В експерименті приймали участь 237 студентів різних курсів та різних спеціальностей галузі 12 Інформаційні технології. В каталог вибіркових дисциплін було включено 54 вибіркових дисципліни з різних освітніх програм та спеціальностей. При розрахунку використовувались статистичні дані з оцінками здобувачів, що вивчали включені в каталог дисципліни раніше та результати опитувань щодо якості викладання навчальних дисциплін, яке проводилось в січні 2023 року.

Результати оцінки роботи рекомендаційної системи в контексті підтримки прийняття рішень при формуванні рекомендації навчальної дисципліни наведено в таблиці 4.1.

Таблиця 4.1 – Результати роботи рекомендаційної системи для формування рекомендації по вибору дисципліни

Показник	Рекомендовано	Не рекомендовано
Обрано	218	19
Не обрано	493	7

Кожному з 237 користувачів було надано по 3 рекомендації дисципліни, таким чином загальна кількість рекомендацій сягнула 711. Користувач міг вибрати з наведених дисциплін лише одну. 7 дисциплін не потрапили до рекомендацій, оскільки 5 з них мали низькі показники рейтингу на основі відгуків та низький середній бал, а 2 дисципліни не відповідали жодному з пошукових запитів і були виключені зі списку при формуванні рекомендації.

218 користувачів погодились з рекомендаціями, наданими системою, та підтвердили бажання вивчати одну із 3 запропонованих дисциплін, записавшись на курс. 19 користувачів обрали дисципліни, що не були запропоновані рекомендаційною системою. Таким чином, показник відгуку склав:

$$\text{Відгук} = \frac{218}{218+19} = 0{,}92.$$

Тобто 92% користувачів погодились з рекомендацією, наданою системою.

Важливим чинником в освітній сфері є не просто забезпечення підтримки вибору, а й задоволеність користувачів зробленим вибором з плином часу. Оцінка задоволеності користувачів зробленим вибором проводилась шляхом інтернет-опитувань протягом 3 тижнів в березні-квітні 2023 року через 1,5-2 місяці після початку вивчення обраних ними дисциплін. Результати опитування серед 218 студентів, вибір яких збігся з рекомендацією, показали наступне:

– 166 студентів в цілому задоволені тим вибором, який вони зробили на основі рекомендації, що становить 76% задоволених користувачів;

– 32 студенти (15%) зазначили, що хотіли б поміняти свій вибір на іншу дисципліну із тих, що пропонувала їм рекомендаційна система;

– 20 студентів (9%) зазначили, що не задоволені вибором.

Аналогічне опитування було проведено серед контрольної групи студентів, які не обирали дисципліни самостійно. В опитуванні прийняло участь 204 студенти, а рекомендації вони отримували самостійно від науково-педагогічних працівників (завідувачі кафедрами, викладачі, що викладають дисципліни вільного вибору), від друзів, на основі інформації на сайті університету та інших джерел:

– 122 студенти в цілому задоволений тим вибором, який вони зробили без використання рекомендаційної системи, що становить 60% задоволених користувачів;

– 47 студентів (23%) зазначили, що хотіли б поміняти свій вибір на іншу дисципліну із тих, що їх також рекомендували друзі, викладачі, чи вони зустрічали опис дисциплін на сайті університету;

– 35 студентів (17%) зазначили, що взагалі не задоволені вибором.

Зведені оцінки наведено на діаграмі на рис. 4.9.

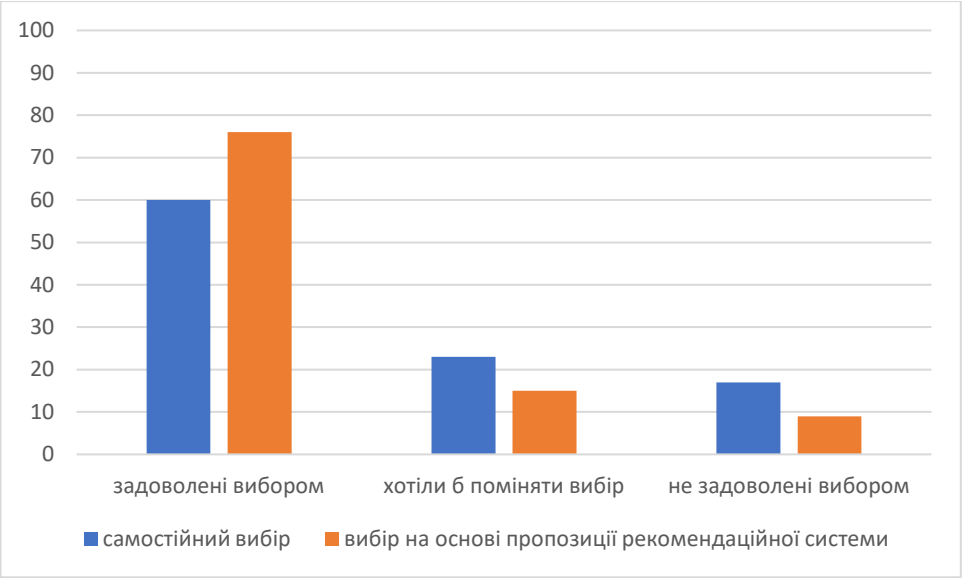


Рис. 4.9. Зведена оцінка якості роботи рекомендаційної системи при виборі дисципліни

Оцінка якості надання рекомендації при вирішенні задачі підбору літературних джерел виконувалась шляхом отримання експертних оцінок щодо релевантності запропонованих рекомендаційною системою варіантів до реальних потреб користувачів. Студентам було запропоновано обов’язково скористатись наданою рекомендацією і після аналізу літературного джерела висловити свою оцінку релевантності за 10-бальною шкалою. В опитуванні прийняло участь 134 студенти. Загальна оцінка рівня релевантності склала в середньому 8,5.

Висновки по розділу 4

В даному розділі наведено приклади практичного застосування розробленої методики побудови рекомендаційної системи для сфери освіти на основі гібридного підходу для деяких типових задач визначення об’єктів інтересу, а саме: рекомендація дисципліни з каталогу вибіркових дисциплін, рекомендація літературного джерела на основі сфери інтересів користувача. Розробка проведена із застосуванням запропонованих в розділі 3 структурної моделі рекомендаційної системи, нечіткої моделі даних та методу побудови рекомендації на основі гібридного підходу з урахуванням характеристик наявних даних про об’єкти інтересу та користувачів системи. Виконано моделювання та проведено оцінку якості наданих рекомендацій. Експериментальним шляхом доведено ефективність використання рекомендаційної системи для підтримки прийняття рішень при виконанні окремих задач освітнього процесу.

ВИСНОВКИ

В роботі вирішено актуальне науково-прикладне завдання підвищення якості надання рекомендацій в сфері освіти за рахунок використання рекомендаційної системи на основі гібридного підходу.

Основні наукові результати, отримані в роботі, можна сформулювати наступним чином:

1. Аналіз моделей та методів обробки даних рекомендаційних систем показав, що окреме використання існуючих методів на основі колаборативної фільтрації та контентної фільтрації не забезпечує необхідних показників якості при наданні рекомендації, а перспективним напрямком є використання гібридних підходів, які, за рахунок комбінації методів, дозволяють уникнути або зменшити проблеми, які виникають при колаборативній та контентній фільтрації. Зокрема, для колаборативних методів серйозну проблему становить збереження даних у вигляді розріджених матриць, а контентні методи мають високу обчислювальну складність та не враховують досвід взаємодії користувача з об’єктами інтересу.

2. Аналіз підходів до розробки більшості рекомендаційних систем та приклади їх реалізації демонструють прив’язку до комерційного аспекту, в той час коли цілі і задачі рекомендаційної системи лежать в іншій площині – рекомендаційна система в сфері освіти повинна підвищити якість освітнього процесу, а не забезпечити зростання прибутку чи розширення клієнтської бази підприємства. Це обумовлює специфічні вимоги до рекомендаційної системи в сфері освіти, в тому числі, пов’язані із, в більшості випадків, разовим застосуванням результатів рекомендації, на відміну від рекомендаційних систем в комерційній, розважальній чи іншій подібній сфері, де одні й ті ж об’єкти інтересу багатократно можуть виступати ціллю для користувача. Ще одним важливим аспектом є те, що організаційні процеси закладу освіти передбачають наявність цільових опитувань серед здобувачів та обов’язкового семестрового оцінювання, що забезпечує наявність інформації про взаємодії здобувачів та дисциплін, здобувачів та викладачів. Це дає змогу зменшити

ефект «холодного старту», на відміну від звичайних рекомендаційних систем, які часто мають подібну проблему.

3. Розроблено модель представлення даних рекомендаційної системи для освітньої сфери на основі нечіткої логіки, яка, на відміну від існуючих, дозволяє використовувати лінгвістичну форму в процесі експертних оцінок параметрів.

4. Розроблено структурну модель рекомендаційної системи для освітньої сфери, яка, на відміну від існуючих, дозволяє формувати комплексну рекомендацію з урахуванням наявних інформаційних ресурсів.

5. Розроблено метод формування рекомендацій на основі гібридного підходу, який, на відміну від існуючих, ґрунтується на нечітких моделях представлення та обробки даних про об’єкти інтересу та користувачів і методах обробки природної мови, що дозволяє підвищити точність рекомендацій.

6. Розроблено методику побудови рекомендаційної системи для сфери освіти на основі гібридного підходу, яка включає структурну модель рекомендаційної системи та метод формування рекомендацій, що ґрунтується на нечітких моделях представлення та обробки даних про об’єкти інтересу та користувачів і методах обробки природньої мови.

7. Враховуючи розроблений алгоритм застосування методики побудови рекомендаційної системи в сфері освіти на основі гібридного підходу було виконано всі необхідні дії для побудови рекомендаційної системи, що вирішує задачі підбору дисципліни вільного вибору та літератури в межах сфери інтересів користувача. Визначено параметри моделей та методів рекомендаційної системи для формування рекомендації навчальної дисципліни та літератури. Досліджено та описано математичні моделі для представлення даних в контексті задачі, що вирішується. Для текстових даних визначено їх характеристики та підібрано необхідні набори методів попередньої обробки. Розроблено структуру баз даних для збереження інформації про об’єкти інтересу (дисципліни та літературні джерела) та

користувачів (студентів). Визначено механізм формування та обробки зовнішніх по відношенню до системи даних, а саме – результатів опитування та семестрових оцінок з дисциплін. Зазначені дії знайшли відображення в програмній реалізації рекомендаційної системи.

6. Апробація рекомендаційної системи проведено для вирішення задач вибору дисципліни блоку вільного вибору та підбору літератури на напрямком інтересів користувача. Застосування запропонованої в дисертаційній роботі методики побудови рекомендаційної системи в сфері освіти на основі гібридного підходу при вирішенні задачі закріплення здобувачів за дисциплінами вільного вибору продемонструвало збіг рекомендації з вибором здобувачів освіти у 92% випадків. При вирішенні задачі підбору літературних джерел оцінка релевантності запропонованих рекомендаційною системою варіантів до реальних потреб користувачів склала в середньому 8,5 за 10-бальною шкалою. Таким чином, наукове завдання виконане, мета роботи досягнута.

Перспективами подальшого розвитку методики є розширення галузі застосування наведених в роботі розробок для побудови рекомендаційних систем не тільки безпосередньо для підтримки освітніх процесів, а й в схожих або дотичних галузях. Зокрема, методика може бути застосована для формування рекомендації щодо вибору виконавця науково-дослідної роботи в науково-дослідних частинах закладів освіти чи науково-дослідних установах, формування пропозицій онлайн-курсів та інших подібних задач, в яких при формуванні рекомендації виникають підзадачі, пов’язані з визначенням подібності текстових наборів даних, та/або потреба обробки даних, представлених у лінгвістичній формі.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Мелешко Є. В. Проблеми сучасних рекомендаційних систем та методи їх рішення / Є. В. Мелешко // Системи управління, навігації та зв'язку. 2018. Вип. 4. С. 120-124. Режим доступу: http://nbuv.gov.ua/UJRN/suntz_2018_4_25

2. Borovyk, Maryna. "Підвищення якості прийняття управлінських рішень за рахунок використання інтелектуальної рекомендаційної системи." *Scientific Notes of Ostroh Academy National University," Economics" Series* 18 (46) (2020): 26-30.

3. Верес, Олег, and Яна Левус. "Рекомендаційна система планування дозвілля в умовах карантину." *Вісник Національного університету «Львівська політехніка». Інформаційні системи та мережі* 11 (2022): 127-144.

4. Паулін, О. М., Г. О. Комлева, and Г. В. Улізко. "Рекомендаційна система для допомоги у вивченні музичних творів." *Вчені записки* (2020): 5202094: 94-99.

5. П’ятикоп О.Є., Мініна К.Г., Безуглов Р.О. Формування рекомендацій фільмів на основі гібридного підходу. Вісник приазовського державного технічного університету. Серія: Технічні науки. Вип. 41. 2020. С 13-19. DOI: 10.31498/2225-6733.41.2020.226119

6. Cherednichenko, O. Y., Ivashchenko, O. V., Gontar, Y. M., & Vorona, B. M. (2018). Інтелектуальний аналіз пропозицій товарів на основі контекстних рекомендацій. *Вісник Національного технічного університету «ХПІ». Серія: Системний аналіз, управління та інформаційні технології*, *1320*(44), 57–66. https://doi.org/10.20998/2079-0023.2018.44.10

7. Zhang, Q., Lu, J., & Zhang, G. (2021). Recommender Systems in E-learning. Journal of Smart Environments and Green Computing, 1(2), 76-89.

8. Omowonuola, Victor & Wilkerson, Bryce & Kher, Shubhalaxmi. (2022). Hybrid Context-Content Based Music Recommendation System. 10.1007/978-3-031-18461-1_8.

9. Patel, Dhruval & Patel, Foram & Chauhan, Uttam. (2023). Recommendation Systems: Types, Applications, and Challenges. 2210-142. 10.12785/ijcds/130168.

10. Meleshko, E. V., S. G. Semenov, and V. D. Khokh. "Дослідження методів побудови рекомендаційних систем в мережі Інтернет." *Системи управління, навігації та зв'язку. Збірник наукових праць* 1.47 (2018): 131-136.

11. Raghuwanshi, Sandeep & Pateriya, R.. (2019). Collaborative Filtering Techniques in Recommendation Systems. 10.1007/978-981-13-6347-4_2.

12. Модель персоналізації рекомендацій контенту на основі колаборативної фільтрації / В. О. Мітрошин, Н. Н. Шаповалова, І. О. Доценко, Н. Х. Саітгареєв // Вісник Криворізького національного університету : зб. наук. праць. – Кривий Ріг, 2021. – Вип. 52. – С. 142–146. – DOI: 10.31721/2306-5451-2021-1-52-142-146. – Бібліогр.: с. 146.

13. I.Mohammed Ali, Sura & Majeed, Sadiq. (2021). A Review of collaborative filtering Recommendation System. Muthanna Journal of Pure Science. 8. 2021. 10.52113/2/08.01.2021/120-131.

14. Yuan, Yiting & Qin, Youyang & Yu, Zekai & Zhang, Congbai. (2023). A Content-based Movie Recommendation System. Theoretical and Natural Science. 2. 56-66. 10.54254/2753-8818/2/20220152.

15. Zhang, Shuting & Liu, Kechen & Yu, Zekai & Feng, Bowen & Ou, Zijie. (2023). Hybrid recommendation system combining collaborative filtering and content-based recommendation with keyword extraction. Applied and Computational Engineering. 2. 927-939. 10.54254/2755-2721/2/20220579.

16. Lytvyn, Vasyl & Vysotska, Victoria & Shatskykh, Viktor & Kohut, Ihor & Petruchenko, Oksana & Dzyubyk, Lyudmyla & Bobrivetc, Vitaliy & Panasyuk, Valentyna & Sachenko, Svitlana & Komar, Myroslav. (2019). Design of a recommendation system based on collaborative filtering and machine learning considering personal needs of the user. Eastern-European Journal of Enterprise Technologies. 4. 6-28. 10.15587/1729-4061.2019.175507.

17. Худік Б.О. Застосування рекомендаційної системи в освітній сфері. Зв'язок, № 6, 2023. С. 58-63. https://doi.org/10.31673/2412-9070.2023.065868

18. Худік Б.О. Модель представлення даних рекомендаційної системи в сфері освіти на основі нечіткої логіки. Том 1 № 21 (2023): Кібербезпека: освіта, наука, техніка. С. 260-272. https://doi.org/10.28925/2663-4023.2032.21.260272

19. Худік Б.О. Метод автоматизованого формування рекомендацій для здобувачів освіти на основі гібридного підходу. Телекомунікаційні та інформаційні технології. Київ, 2022, №4 (81). С. 21-27. https://doi.org/10.31673/2412-4338.2023.042127

20. Bohdan Khudik. Analysis of the approaches to the development of the multiagent recommender system based on the fuzzy logic// ITU Workshop for Europe and CIS Regions. ICT infrastructure as a basis for digital economy. 14-16 may 2019. International Telecommunication Union, State University of Telecommunications. P. 56.

21. Худік Б.О. Визначення рейтингу навчальної дисципліни на основі інтернет-опитувань // Міжнародна науково-практична конференція «Сучасні аспекти діджиталізації та інформатизації в програмній та комп'ютерній інженерії». Збірник тез. К.: ДУТ, 2023. С. 167-169.

22. Різанова С.Д., Худік Б.О. Інтеграція штучного інтелекту для оптимізації процесів розробки веб-дизайну, інтерфейсів та супутніх аспектів в цифровому просторі// Всеукраїнська науково-технічна конференція «Технологічні горизонти: дослідження та застосування інформаційних технологій для технологічного прогресу України і Світу». Збірник тез. К.: ДУІКТ, 2023. С. 160-161.

23. Худік Б.О. Використання нечітких моделей даних в рекомендаційних системах// Науково-практична конференція «Проблеми комп'ютерної інженерії». Збірник тез. К.: ДУІКТ, 2023. С. 154-155.

24. Худік Б.О. Вилучення інформації з освітніх джерел на основі Natural Language Processing// «ТАК»: телекомунікації, автоматика, комп'ютерно-інтегровані технології: зб. доповідей Всеукр. наук.-практ. конф. молодих вчених, 5-6 грудня 2023 р./ ДВНЗ «ДонНТУ»; відп. ред. Г.В. Ступак. Луцьк: ДВНЗ «ДонНТУ», 2023. С. 27-29.

25. Yadav N., Kumar R., Singh A., Pal S. Diversity in Recommendation System: Cluster Based Approach. Hybrid Intelligent Systems. 2020. 113–122c.

26. John S. Breese, David Heckerman, and Carl Myers Kadie. Empirical analysis of predictive algorithms for collaborative filtering. CoRR, abs/1301.7363, 2013. 69c.

27. Liu Na, Ming-Xia Li, Qiu Hai-yang, and Hao-Long Su. A hybrid user-based collaborative filtering algorithm with topic model. Appl. Intell., 51(11):7946–7959, 2021.

28. Soumen Chakrabarti. Mining the web - discovering knowledge from hypertext data. Morgan Kaufmann, 2003.

29. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. Cambridge University Press, 2008.

30. Yeruva, Sagar & Sathvika, Addi & Sruthi, Damera & Reddy, Duggasani & Krishna, Gogineni. (2022). Apparel Recommendation System using Content-Based Filtering. International Journal of Recent Technology and Engineering (IJRTE). 11. 46-51. 10.35940/ijrte.D7331.1111422.

31. Penugonda, Anusha & Pantula, Muralidhar. (2023). Assessment on Multiple Aspects of Online Book Recommendation Systems. 10.1007/978-981-19-8497-6_57.

32. Robin D. Burke. Hybrid recommender systems: Survey and experiments. User Model. User Adapt. Interact., 12(4):331–370, 2002.

33. Beheshti, A., Yakhchi, S., Mousaeirad, S., Ghafari, S. M., Goluguri, S. R., & Edrisi, M. A. (2020). Towards cognitive recommender systems. Algorithms, 13(8), 176.

34. Raghuwanshi, Sandeep & Pateriya, R.. (2019). Collaborative Filtering Techniques in Recommendation Systems. 10.1007/978-981-13-6347-4_2.

35. Yuan, Zhenning & Lee, Jong & Zhang, Sai. (2021). Optimization of the Hybrid Movie Recommendation System Based on Weighted Classification and User Collaborative Filtering Algorithm. Complexity. 2021. 1-13. 10.1155/2021/4476560.

36. Puri, Gaurica & Varanasi, Abhiram & Singh, Gagandeep & Agarwal, Harshit & Tomar, Ravi & Choudhury, Tanupriya. (2023). Investigating Graph-Based Recommendations Systems and Graph Traversal Algorithms. 10.1007/978-981-19-9512-5_11.

37. Jonathan L. Herlocker, Joseph A. Konstan, John Riedl (2000). Explaining collaborative filtering recommendations. CSCW '00: Proceedings of the 2000 ACM conference on Computer supported cooperative work. Pages 241–250 https://doi.org/10.1145/358916.358995

38. Billsus D., Pazzani M.J. Content-Based Recommendation Systems (1998). The Adaptive Web. Vollume 4321. Pages 325-341.

39. Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2018. Deep Learning based Recommender System: A Survey and New Perspectives. ACM Comput. Surv. 1, 1, Article 1 (July 2018), 35 pages. DOI: 0000001.0000001

40. Brin S. An Empirical Comparison of Techniques for Extracting Structure from Texts (1999). Early visual learning. Pages 161–181.

41. M. D. Ekstrand, J. T. Riedl, J. A. Konstan, Collaborative filtering recommender systems, Found. Trends Hum.- Comput. Interact. 4 (2) (2011) 81–173. doi:10.1561/1100000009.

42. R. Burke, Knowledge-based recommender systems, in: A. Kent (Ed.), Encyclopedia of Library and Information Science, Vol. 69, CRC Press, 2000, pp. 181–201.

43. A.Felfernig,R.Burke,Constraint based recommender systems: Technologies and research issues, in:Proceedings of the 10th International Conference on Electronic Commerce, ICEC ’08, ACM, New York, NY, USA, 2008, pp. 3:1–3:10. doi:10.1145/1409540.1409544

44. M. Balabanovic´, Y. Shoham, Fab: Content-based, collaborative recommendation, Commun. ACM 40 (3) (1997) 66–72. doi:10.1145/245108.245124.

45. B. M. Sarwar, J. A. Konstan, A. Borchers, J. Herlocker, B. Miller, J. Riedl, Using filtering agents to improve prediction quality in the grouplens research collaborative filtering system, in: Proceedings of the 1998 ACM Con- ference on Computer Supported Cooperative Work, CSCW ’98, ACM, New York, NY, USA, 1998, pp. 345–354. doi:10.1145/289444.289509

46. Liu L., Mehandjiev N., Xu D.L. Multi-Criteria Service Recommendation Based on User Criteria Preferences. (2011). 5th ACM Conference on Recommender Systems, Chicago, Oct. 23rd-27th.

47. Son L.H.? HU-FCF: A hybrid user-based fuzzy collaborative filtering method in Fecommender Systems. International Journal of Expert Systems with Applications 41 (2014) 68616870.

48. Choi S.H., Jeong Y.S., Jeong M.K. A Hybrid Recommendation Method with Reduced Data for Large-Scale Applicationn. IEEE Transactions on systems, man and cybernetics – Part C: Applications and Reviews, VOL.40, No.5, September 2010.

49. Kardan A.A., Ebrahimi M. A novel approach to hybrid recommendation systems based on association rules mining for content recommendation in asunchronous discussion groups. International Journal of Information Science 219 (2013) 93110.

50. Christensen I., Schiaffino S. A Hybrid Approach for Group Profiling in Recommender Systems. Journal of Universal Computer Science, vol. 20, no.4, 2014.

51. Zheng X., Ding W., Xu J., Chen D. Personalized recommendation based on review topics. SOCA (2014) 8:1531 DOI: 10.1007/s11761-013-0140-8.

52. Lekakos G., Caravelas P. A hybrid approach for improving predictive accuracy of collaborative filtering algorithms. User Model User-Adap Anter (2007) 17:540. DOI:10.1007/s11257-006-9019-0, Springer.

53. Baranchuk, Nina & Prasad, Ashutosh. (2022). Design of Product Quality Scales for Conveying Information by Infomediaries. International Journal of Research in Marketing. 10.1016/j.ijresmar.2022.07.002

54. Tartaraj, Azeta & Trebicka, Brunela. (2023). Optimal Grading Scales for Enhancing Product Evaluation by Infomediaries. Interdisciplinary Journal of Research and Development. 10. 59. 10.56345/ijrdv10n109.

55. Редьога Ю.В. Арифметизація ординальних шкал вимірювання якості програмних засобів / Ю.В. Редьога, Н.А. Яремчук // Інформаційні системи, механіка та керування : наук.- техн. зб. – 2011. – Вип. 7. – С. 5-15.

56. Kostenko, Iarislav, and Mykola Sydorov. "Аналіз рангових шкал у масових опитуваннях." *Вісник ХНУ імені ВН Каразіна. Серія «Соціологічні дослідження сучасного суспільства: методологія, теорія, методи»* 39 (2017): 110-117.

57. Нескородєва, І. І., and С. А. Пустовгар. "Побудова інтервальної шкали оцінювання рівня фінансового ризику підприємств машинобудівної галузі." (2018).

58. Яремчук, Н. А., and Р. С. Семенюк. "Способи урахування невизначеності при побудові лінгвістичних шкал." *Український метрологічний журнал/Ukrainian Metrological Journal* 3А (2020): 77-81.

59. Матвіїшина Н. В., Бровченко Є. Е. Класифікація шкал вимірювання оцінки знань. Вісник Запорізького державного університету. Фізикоматематичні науки. Біологічні науки. 1999. № 2. С. 69-73.

60. Григорук П.М. Шкалювання в маркетингових дослідженнях / П.М. Григорук // Вісник Хмельницького національного університету, 2009. – .№5.- с.138-142. – (Серія «Економічні науки», т.3)

61. Зуєва, Олександра Валерівна. "Формування шкали оцінювання пріоритетності складових фінансового потенціалу банку." *Науковий вісник Ужгородського університету. Серія «Економіка»* 1.49 (2017): 258-263.

62. Aleksashyn, V. S., and H. Y. Mykhalchuk. "Метод ранжування об'єктів для пошуку за ключовими словами у децентралізованій мережі." *Актуальні проблеми автоматизації та інформаційних технологій* 23 (2019).

63. Главчева, Юлія Миколаївна, et al. "Розробка підходу для ранжування академічних установ за показниками наукової діяльності." *Сучасні інформаційні системи* 3,№ 1 (2019): 63-70.

64. Куртов, А. І., et al. "Експертні оцінки. Метод Делфі як технологія прийняття управлінських рішень." *Збірник наукових праць Харківського університету Повітряних Сил* 1 (2017): 118-122.

65. Турчина, В. А., and В. О. Гринченко. "Порівняльна характеристика основних методів експертних оцінок." *Міжнародний науковий комітет* (2020): 252.

66. Ni, Ansong & Azerbayev, Zhangir & Mutuma, Mutethia & Feng, Troy & Zhang, Yusen & Yu, Tao & Hassan, Ahmed & Radev, Dragomir. (2021). SummerTime: Text Summarization Toolkit for Non-experts.

67. Xiao, Hui & Zhang, Yao & Kou, Gang & Zhang, Si & Branke, Juergen. (2023). Ranking and selection for pairwise comparison. Naval Research Logistics (NRL). 70. 10.1002/nav.22093.

68. Синенко М.А. Метод Сааті при прийнятті управлінських рішень на прикладі підприємства малого бізнесу. Бізнес та інтелектуальний капітал. Інтелект XXI № 1 ‘2018. С. 235-238.

69. Saaty Thomas L. Mathematical Principles of Decision Making: (Principia Mathematica Decernendi) RWS Publications (2009) ISBN 10: 1888603100 ISBN 13: 9781888603101

70. Saaty T. Decision Making with the Analytic Network Process / T. Saaty, L. Vargas. – New York: Springer, 2013. – 363 с. – (Springer Science+Business 64 Media). – (International Series in Operations Research & Management Science; 195).

71. Luce, Christine & Kirnan, Jean. (2016). Using Indirect vs. Direct Measures in the Summative Assessment of Student Learning in Higher Education.

Journal of the Scholarship of Teaching and Learning. 16. 75. 10.14434/josotl.v16i4.19371.

72. Гнатієнко Г.М., Снитюк В.Є. Експертні технології прийняття рішень: Монографія. – К.: ТОВ «Маклаут», 2008. – 444 с.

73. Грабовецький Б. Методи експертних оцінок: теорія, методологія, напрямки використання:[монографія] / Б.Є. Грабовецький. – Вінниця: ВНТУ, 2010. - 171 с.

74. Волошин О.Ф., Гнатієнко Г.М. Проблеми узгодженості експертної інформації в задачах прийняття рішень // Знання-Діалог-Рішення, KDS-97: Збірник наукю праць Міжн. конф. – Ялта 1997.

75. Рудень В. Методика проведення та оцінки результатів експертних оцінок / Рудень В., Гутор Т. //Український медичний часопис. - 2011. - №2 (82). -С. III – IV.

76. Новосад В. П. Оцінювання узгодженості експертних висновків методами рангової кореляції / В. П. Новосад // Демократичне врядування в контексті глобальних викликів та кризових ситуацій : матеріали наук.-практ. конф. за міжнар. участю (3 квіт. 2009 р., м. Львів) : у 2 ч. - Ч. 2 / за наук. ред. чл.-кор. НАН України В. С. Загорського, доц. А. В. Ліпенцева. - Львів : ЛРІДУ НАДУ, 2009. - С. 353-356.

77. Everitt BS and Hothorn T (2009). A Handbook of Statistical Analyses Using R, 2nd ed. Chapman and Hall, London.

78. Kendall, M.G. (1975) Rank Correlation Methods. 4th Edition, Charles Griffin, London.

79. L. A. Zadeh. Fuzzy Sets. *Information and control*. 8, (1965): 338-353.

80. Bede, Barnabas, and Barnabas Bede. Fuzzy Analysis. Springer Berlin Heidelberg, 2013.

81. Deng, Jixiang, and Yong Deng. "Information volume of fuzzy membership function." International Journal of Computers Communications & Control 16.1 (2021).

82. Штовба С.Д. Побудова функції належності нечітких множин за кластеризацією експериментальних даних. Інформаційні технології та комп’ютерна інженерія. 2006. №2. С. 92–95.

83. Fuzzy Logic Toolbox. User’s Guide. The MathWorks, Inc., 1999. – 134 p

84. Kaufmann, A., Gupta, M. : Introduction to Fuzzy Arithmetic: Theory and Applications. Van Nostrand Reinhold Company, New York (1985).

85. Джеджула, в., & Єпіфанова, і. (2021). Методичні підходи до побудови функцій належності нечітких математичних моделей. Modeling the development of the economic systems, (1), 42–46.

86. О. Архипов Є., С. Архипова. А. Оцінювання якості роботи експертів за даними багатооб’єктної експертизи // Захист інформації. — 2011. — №4(53) — С. 45-54.

87. Capocelli R. Fuzzy sets and decision theory. R. Capocelli, A. De Luca — Information and Control. – 2003 - v. 23 - P. 446—473.

88. Id, & Trillas, Enric & de Soto, Adolfo & Tfss, Transactions On Fuzzy Sets And Systems. (2022). On a New View of a Fuzzy Set. Transactions on Fuzzy Sets and Systems. 2. 10.30495/tfss.2022.1971064.1051.

89. Jiang, Lisheng. (2020). Double-quantified linguistic variable. Information Sciences. 545. 10.1016/j.ins.2020.08.026.

90. Zolotukhina O. A. Using fuzzy sets to unify the presentation of some types of imperfect information in resource supervising systems. *Sciences of Europe*. 2018. Vol. 2. No. 27. P. 55-59.

91. Джеджула, Вячеслав, and Ірина Єпіфанова. "Методичні підходи до побудови функцій належності нечітких математичних моделей." *Modeling the development of the economic systems* 1 (2021): 42-46.

92. Gu, Suhang & Chou, Yongxin & Zhou, Jie & Jiang, Zhibin & Lu, Mingli. (2023). Takagi–Sugeno–Kang Fuzzy Clustering by Direct Fuzzy Inference on Fuzzy Rules. IEEE Transactions on Emerging Topics in Computational Intelligence. PP. 1-16. 10.1109/TETCI.2023.3336537.

93. Benferhat S., Dubois D., Prade H. (1992) Representing defaultrules in possibilistic logic. Proc. of the 3rd Inter. Conf. on Principles of Knowledge Representation and Reasoning(KR'92) (B. Nebel, C. Rich, W. Swartout, eds.), Cambridge, MA, Oct. 25-29, 673-684.7

94. Dubois D., Prade H. (1989) A typology of fuzzy "if... then..." rules. Proc. of the 3rd Inter.Fuzzy Systems Assoc. (IFSA) Congress, Seattle, WA, Aug. 6-11, 782-785.

95. Piegat, Andrzej. (2001). Fuzzy Modeling and Control. 10.1007/978-3-7908-1824-6.

96. Liu, J. Intelligent Control Design and Matlab Simulation. — Singapore: Springer, 2017. — 294 p.

97. Mohinder Jankiraman, Space Time Codes and MIMO Systems, Universal personal communication, Aston house, Boston, London, 2004.

98. Mamdani, E. H., Application of fuzzy algorithms for the control of a simple dynamic plant. In Proc IEEE (1974), 121—159.

99. Adriyendi. (2018). Fuzzy Logic using Tsukamoto Model and Sugeno Model on Prediction Cost. International Journal of Intelligent Systems and Applications. 6. 10.5815/ijisa.2018.06.02.

100. Geetha, Dr & Gomathy, Dr & Yagn, Mr.D.Sri & Praneesh, Sai. (2023). The role of natural language processing. *Interantional journal of scientific research in engineering and management*. 07. 1-11. 10.55041/IJSREM27094.

101. Geetha, Dr & Gomathy, Dr & Ram, Mr.P.V. & N, Surya. (2023). Novel study on natural language processing. *Interantional journal of scientific research in engineering and management.* 07. 1-11. 10.55041/IJSREM27091.

102. Chowdhary, KR1442, and K. R. Chowdhary. "Natural language processing." *Fundamentals of artificial intelligence* (2020): 603-649.

103. Sharp, Bernadette. (2015). Towards a Cognitive Natural Language Processing Perspective. 10.1007/978-3-319-08043-7_3.

104. Chakraborty, Sutapa. (2023). Sentiment Analysis in the Perspective of Natural Language Processing. *International Journal for Research in Applied Science and Engineering Technology*. 11. 2235-2241. 10.22214/ijraset.2023.56925.

105. Gavrylenko, S. & Zozulia, V. (2022). Дослідження методів виявлення аномалій на етапі попередньої обробки даних. Системи управління, навігації та зв’язку. Збірник наукових праць. 1. 52-56. 10.26906/SUNZ.2022.1.052.

106. Якименко, Дмитро & Катаєва, Євгенія. (2022). Методи та засоби інтелектуального аналізу текстових документів. Вісник Черкаського державного технологічного університету. 43-52. 10.24025/2306-4412.2.2022.259408.

107. Tabassum, Ayisha, and Rajendra R. Patil. "A survey on text pre-processing & feature extraction techniques in natural language processing." *International Research Journal of Engineering and Technology (IRJET)* 7.06 (2020): 4864-4867.

108. Sarica, Serhad, and Jianxi Luo. "Stopwords in technical language processing." Plos one 16.8 (2021): e0254937.

109. Kaur, Jashanjot, and Preetpal Kaur Buttar. "Stopwords removal and its algorithms based on different methods." *International Journal of Advanced Research in Computer Science* 9.5 (2018): 81-88.

110. Aliero, Abubakar Ahmad, et al. "Systematic Review on Text Normalization Techniques and its Approach to Non-Standard Words." *International Journal of Computer Applications* 975: 8887.

111. Karaa, Wahiba Ben Abdessalem. "A new stemmer to improve information retrieval." *International Journal of Network Security & Its Applications* 5.4 (2013): 143.

112. Polus, Manhal Elias, and Thekra Abbas. "Development for performance of Porter Stemmer algorithm." *Eastern-European Journal of Enterprise Technologies* 1.2 (2021): 109.

113. Balakrishnan, Vimala, and Ethel Lloyd-Yemoh. "Stemming and lemmatization: A comparison of retrieval performances." (2014): 174-179.

114. Khyani, Divya, et al. "An interpretation of lemmatization and stemming in natural language processing." *Journal of University of Shanghai for Science and Technology* 22.10 (2021): 350-357.

115. Üstün, Ahmet, Murathan Kurfalı, and Burcu Can. "Characters or morphemes: How to represent words?." *Association for Computational Linguistics*, 2018.

116. Creutz, Mathias, and Krista Lagus. "Unsupervised models for morpheme segmentation and morphology learning." *ACM Transactions on Speech and Language Processing (TSLP)* 4.1 (2007): 1-34.

117. Hapke, Hannes, Cole Howard, and Hobson Lane. *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python*. Simon and Schuster, 2019.

118. Millstein, Frank. *Natural language processing with python: natural language processing using NLTK*. Frank Millstein, 2020.

119. Reese, Richard M., and AshishSingh Bhatia. *Natural Language Processing with Java: Techniques for building machine learning and neural network models for NLP*. Packt Publishing Ltd, 2018.

120. Qader, Wisam A., Musa M. Ameen, and Bilal I. Ahmed. "An overview of bag of words; importance, implementation, applications, and challenges." *2019 international engineering conference (IEC)*. IEEE, 2019.

121. HaCohen-Kerner, Yaakov, Daniel Miller, and Yair Yigal. "The influence of preprocessing on text classification using a bag-of-words representation." *PloS one* 15.5 (2020): e0232525.

122. Walkowiak, Tomasz, Szymon Datko, and Henryk Maciejewski. "Bag-of-words, bag-of-topics and word-to-vec based subject classification of text documents in polish-a comparative study." *Contemporary Complex Systems and Their Dependability: Proceedings of the Thirteenth International Conference on*

Dependability and Complex Systems DepCoS-RELCOMEX, July 2-6, 2018, Brunów, Poland 13. Springer International Publishing, 2019.

123. Kim, Sang-Woon, and Joon-Min Gil. "Research paper classification systems based on TF-IDF and LDA schemes." *Human-centric Computing and Information Sciences* 9 (2019): 1-21.

124. Jalilifard, Amir, et al. "Semantic sensitive TF-IDF to determine word relevance in documents." *Advances in Computing and Network Communications: Proceedings of CoCoNet 2020, Volume 2*. Singapore: Springer Singapore, 2021. 327-337.

125. Cahyani, Denis Eka, and Irene Patasik. "Performance comparison of tf-idf and word2vec models for emotion text classification." *Bulletin of Electrical Engineering and Informatics* 10.5 (2021): 2780-2788.

126. Barman, Dibyendu, and Md Bakash Ahamed. "Improved LZW Compression Technique using Difference Method." *International Journal of Innovative Technology and Exploring Engineering* 9.5 (2020): 87-92.

127. Hon, Wing-Kai & Ku, Tsung & Shah, Rahul & Thankachan, Sharma. (2013). Space-Efficient Construction Algorithm for the Circular Suffix Tree. Proceedings of the Data Compression Conference. 7922. 496-496. 10.1109/DCC.2013.76.

128. Pibiri, Giulio Ermanno, and Rossano Venturini. "Techniques for inverted index compression." *ACM Computing Surveys (CSUR)* 53.6 (2020): 1-36.

129. Janani, R., and S. Vijayarani. "Text document clustering using spectral clustering algorithm with particle swarm optimization." *Expert Systems with Applications* 134 (2019): 192-200.

130. Ali, Zuhair Hussein, Amal Abbas Kadhim, and Azal Minshed Abid. "Subject review: text clustering algorithms." *International Journal of Engineering Research and Advanced Technology (IJERAT)* 6.9 (2020).

131. Juan Luis Suárez, Salvador García, Francisco Herrera,A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental

analysis, prospects and challenges, Neurocomputing,Volume 425, 2021, Pages 300-322, ISSN 0925-2312

132. Baturinets A. (2022) Distance measures-based information technology for identifying similar data series. Scientific Journal of TNTU (Tern.), vol. 105, no 1, pp. 128-140.

133. Park, Kwangil, June Seok Hong, and Wooju Kim. "A methodology combining cosine similarity with classifier for text classification." *Applied Artificial Intelligence* 34.5 (2020): 396-411.

134. Wang, Jiapeng, and Yihong Dong. "Measurement of text similarity: a survey." *Information* 11.9 (2020): 421.

135. Nasar, Zara, Syed Waqar Jaffry, and Muhammad Kamran Malik. "Named entity recognition and relation extraction: State-of-the-art." *ACM Computing Surveys (CSUR)* 54.1 (2021): 1-39.

136. Al-Moslmi, Tareq, et al. "Named entity extraction for knowledge graphs: A literature overview." *IEEE Access* 8 (2020): 32862-32881.

137. Shelar, Hemlata, et al. "Named entity recognition approaches and their comparison for custom ner model." *Science & Technology Libraries* 39.3 (2020): 324-337.

138. Pressé, Steve & Sgouralis, Ioannis. (2023). Hidden Markov Models. 10.1017/9781009089555.011.

139. Angulo, J. & Ruiz-Medina, María. (2023). Informational assessment of large scale self-similarity in nonlinear random field models. Stochastic Environmental Research and Risk Assessment. 1-15. 10.1007/s00477-023-02541-x.

140. Naseer, Salman, et al. "Named Entity Recognition (NER) in NLP Techniques, Tools Accuracy and Performance." *Pakistan Journal of Multidisciplinary Research* 2.2 (2021): 293-308.

141. Moon, Seonghyeon & Chung, Sehwan & Chi, Seokho. (2020). Bridge Damage Recognition from Inspection Reports Using NER Based on Recurrent Neural Network with Active Learning. Journal of Performance of Constructed Facilities. 34. 04020119. 10.1061/(ASCE)CF.1943-5509.0001530.

142. Patil, Archana & Ghumbre, Shashikant & Attar, Vahida. (2023). Named Entity Recognition over Dialog Dataset Using Pre-trained Transformers. 10.1007/978-981-99-1414-2_43.

143. Wolf, Thomas & Debut, Lysandre & Sanh, Victor & Chaumond, Julien & Delangue, Clement & Moi, Anthony & Cistac, Pierric & Rault, Tim & Louf, Remi & Funtowicz, Morgan & Davison, Joe & Shleifer, Sam & Platen, Patrick & Ma, Clara & Jernite, Yacine & Plu, Julien & Xu, Canwen & Scao, Teven & Gugger, Sylvain & Rush, Alexander. (2020). Transformers: State-of-the-Art Natural Language Processing. 38-45. 10.18653/v1/2020.emnlp-demos.6.

144. Du, Qiming & Li, Nan & Liu, Wenfu & Sun, Daozhu & Yang, Shudan & Yue, Feng. (2022). A Topic Recognition Method of News Text Based on Word Embedding Enhancement. Computational Intelligence and Neuroscience. 2022. 1-15. 10.1155/2022/4582480.

145. Yin W., Kann K., Yu M., Schütze H. Comparative study of cnn and rnn for natural language processing. – Berlin: Connel, 2017. – pp. 126.

146. Suzuki, Masayuki, et al. "Improvements to n-gram language model using text generated from neural language model." *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

147. Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Jauvin, A Neural Probabilistic Language Model (2003), Journal of Machine Learning Research. https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf

148. Mikolov, Tomas & Chen, Kai & Corrado, G.s & Dean, Jeffrey. (2013). Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR. 2013.

149. Goldberg, Yoav & Levy, Omer. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method.

150. M. Grbovic & H. Cheng, Real-time personalization using embeddings for search ranking at Airbnb, 2018.

151. J. Wang et. al., Billion-scale commodity embedding for e-commerce recommendation in Alibaba, 2018

152. Sun, Weikai & ma, Yinghua & Li, Shenghong & Zhang, Shiyi. (2020). Similar Cluster Based Continuous Bag-of-Words for Word Vector Training. 10.1007/978-981-13-9409-6_106.

153. Tang, Yachun. (2022). Research on Word Vector Training Method Based on Improved Skip-Gram Algorithm. Advances in Multimedia. 2022. 1-8. 10.1155/2022/4414207.

154. Haider, Mofiz Mojib, et al. "Automatic text summarization using gensim word2vec and k-means clustering algorithm." *2020 IEEE Region 10 Symposium (TENSYMP)*. IEEE, 2020.

155. Mohammed, Shapol M., Karwan Jacksi, and Subhi RM Zeebaree. "Glove word embedding and DBSCAN algorithms for semantic document clustering." *2020 International Conference on Advanced Science and Engineering (ICOASE)*. IEEE, 2020.

156. Yao, Tengjun, Zhengang Zhai, and Bingtao Gao. "Text classification model based on fasttext." *2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS)*. IEEE, 2020.

157. Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

158. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

159. Koroteev, M. V. "BERT: a review of applications in natural language processing and understanding." *arXiv preprint arXiv:2103.11943* (2021).

160. Gillioz, Anthony, et al. "Overview of the Transformer-based Models for NLP Tasks." *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2020.

161. Bhola, Akshay, et al. "Retrieving skills from job descriptions: A language model based extreme multi-label classification framework." *Proceedings of the 28th international conference on computational linguistics*. 2020.

162. Ткаченко, Л. В., & Хмельницька, О. С. (2021). Особливості впровадження дистанційного навчання в освітній процес закладу вищої освіти. Педагогіка формування творчої особистості у вищій і загальноосвітній школах. № 75, Т. 3. С.91-96. Doi: https://doi.org/10.32840/1992-5786.2021.75-3.18

163. Нікіфорова, Т. Д., Капшук, О. А., & Нечепуренко, Д. (2022). Навчальна та організаційна робота деканату в умовах воєнного стану. Матеріали ІІ Форуму академічної спільноти «Освіта в умовах війни: реалії, виклики та шляхи подолання». С.41-43.

164. Аліпа, Михайло, Артем Куриленко, and Сергій Волощук. "Jцінка ефективності рекомендаційних систем IPTV, що врахують неявні уподобання користувачів." *Наука та виробництво* 23 (2020): 193-202.

165. Mena-Maldonado, Elisa, et al. "Agreement and disagreement between true and false-positive metrics in recommender systems evaluation." *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020.

166. Mena-Maldonado, Elisa, et al. "Popularity bias in false-positive metrics for recommender systems evaluation." *ACM Transactions on Information Systems (TOIS)* 39.3 (2021): 1-43.

ДОДАТОК А.

СПИСОК ПУБЛІКАЦІЙ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Наукові праці в яких опубліковані основні наукові результати дисертації:

1. Худік Б.О. Застосування рекомендаційної системи в освітній сфері. Зв'язок, № 6, 2023, С. 47-52.
2. Худік Б.О. Модель представлення даних рекомендаційної системи в сфері освіти на основі нечіткої логіки. Том 1 № 21 (2023): Кібербезпека: освіта, наука, техніка. С. 260-272. https://csecurity.kubg.edu.ua/index.php/journal/article/view/530
3. Худік Б.О. Метод автоматизованого формування рекомендацій для здобувачів освіти на основі гібридного підходу. Телекомунікаційні та інформаційні технології. Київ, 2022, №4 (81), С. 21-27. https://tit.dut.edu.ua/index.php/telecommunication/article/view/2490

Наукові праці, які засвідчують апробацію матеріалів дисертації:

4. Bohdan Khudik. Analysis of the approaches to the development of the multiagent recommender system based on the fuzzy logic// ITU Workshop for Europe and CIS Regions. ICT infrastructure as a basis for digital economy. 14-16 may 2019. International Telecommunication Union, State University of Telecommunications. P. 56.
5. Худік Б.О. Визначення рейтингу навчальної дисципліни на основі інтернет-опитувань // Міжнародна науково-практична конференція «Сучасні аспекти діджиталізації та інформатизації в програмній та комп'ютерній інженерії». Збірник тез. К.: ДУТ, 2023. С. 167-169.
6. Різанова С.Д., Худік Б.О. Інтеграція штучного інтелекту для оптимізації процесів розробки веб-дизайну, інтерфейсів та супутніх аспектів в цифровому просторі// Всеукраїнська науково-технічна конференція «Технологічні горизонти: дослідження та застосування інформаційних

технологій для технологічного прогресу України і Світу». Збірник тез. К.: ДУІКТ, 2023. С. 160-161.

7. Худік Б.О. Використання нечітких моделей даних в рекомендаційних системах// Науково-практична конференція «Проблеми комп'ютерної інженерії». Збірник тез. К.: ДУІКТ, 2023. С. 154-155.

8. Худік Б.О. Вилучення інформації з освітніх джерел на основі Natural Language Processing// «ТАК»: телекомунікації, автоматика, комп'ютерно-інтегровані технології: зб. доповідей Всеукр. наук.-практ. конф. молодих вчених, 5-6 грудня 2023 р./ ДВНЗ «ДонНТУ»; відп. ред. Г.В. Ступак. Луцьк: ДВНЗ «ДонНТУ», 2023. С. 27-29.

Документ підписано у сервісі Вчасно (початок)
Khudik-PHD-1-162.pdf

Документ підписано у сервісі Вчасно (продовження)
Khudik-PHD-1-162.pdf

Документ відправлено: 11:33 27.12.2023

Власник документу

Електронний підпис
11:33 27.12.2023
Ідентифікаційний код: 3437612915
ХУДІК БОГДАН ОЛЕКСАНДРОВИЧ
Власник ключа: ХУДІК БОГДАН ОЛЕКСАНДРОВИЧ
Час перевірки КЕП/ЕЦП: 11:33 27.12.2023
Статус перевірки сертифікату: Сертифікат діє
Серійний номер: 5E984D526F82F38F0400000024BED0000B22AC04
Тип підпису: удосконалений