Міністерство освіти і науки україни
Національний університет «Львівська політехніка»

Кваліфікаційна наукова
праця на правах рукопису

Шварц Михайло Євгенійович

УДК 004.89

ДИСЕРТАЦІЯ
ГІБРИДНІ МОДЕЛІ І МЕТОДИ ПРОГНОЗУВАННЯ РЕКОМЕНДАЦІЙ ДЛЯ ІНТЕРНЕТ-МАГАЗИНУ.

01.05.03 – Математичне та програмне забезпечення обчислювальних машин і систем
05 «Технічні науки»

Подається на здобуття наукового ступеня кандидата технічних наук

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

Науковий керівник:
Лобур Михайло Васильович,
д.т.н., професор

Ідентичність всіх примірників дисертації
ЗАСВІДЧУЮ:
Вчений секретар спеціалізованої
вченої ради Д 35.052.05 Р.А. Бунь

Львів – 2019

АНОТАЦІЯ

Шварц М.Є. Гібридні моделі і методи прогнозування рекомендацій для інтернет-магазину. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук (доктора філософії) за спеціальністю 01.05.03 – «Математичне і програмне забезпечення обчислювальних машин і систем» – Національний університет «Львівська політехніка» , Міністерство освіти і науки України, Львів, 2019.

Зміст дисертації. У дисертаційній роботі розв'язано наукове завдання розроблення і дослідження гібридних моделей і методів прогнозування рекомендацій для інтернет-магазину, які можуть бути використані і при функціонуванні інших суб'єктів електронної комерції, таких, як електронні торгові ряди, інтернет-вітрини, інформаційно-пошукові системи, пошукові системи в глобальній мережі Інтернет. Основне призначення рекомендаційних систем – надання рекомендацій користувачам при виборі предметів, які найбільше відповідають їх інтересам та вподобанням. Предметами можуть бути товари, об'єкти або послуги.

Розробленню і впровадженню рекомендаційних систем сприяв стрімкий розвиток інформаційно-комунікаційних технологій, а саме, Інтернет, як всесвітньої мережі для зберігання і передачі інформації, а також Всесвітньої павутини (WWW, Word Wide Web). На даний час в Інтернеті зберігається $1{,}2 \cdot 10^{9}$ сайтів, які містять $16{,}2 \cdot 10^{70}$ байтів інформації. Прогнозується, що до 2020 року цей показник зросте до $44 \cdot 10^{70}$ байтів. Таким чином на даний час проблема полягає не у відсутності інформації, а у відсутності ефективних механізмів пошуку інформації. Сучасні пошукові системи (Google, Yahoo) повертають значно більший об'єм інформації, ніж користувач здатний обробити. Користувачеві може не вистачати знань, часу або досвіду, або того й іншого, щоб вибрати те, що відповідає його потребам. Користувач явно або побічно надає системі інформацію про свої уподобання. Таким чином, рекомендаційна система для інтернет-магазину подається у вигляді системи (програми), що використовує певний алгоритм фільтрації та наявну

інформацію про потреби користувача, щоб рекомендувати йому набір об'єктів, які він вважає найбільш корисними для себе.

У дисертаційній роботі отримав подальший розвиток метод розрахунку коефіцієнтів подібності векторів профілів користувачів і векторів профілів предметів, який на відміну від існуючих, використовує демографічні характеристики користувачів, що дозволяє підвищити точність прогнозування рекомендацій і визначати коефіцієнти подібності для нового користувача і нового предмета.

На основі концепції застосування в одному методі категоріальної, мішаної і числової кластеризації вперше розроблено метод пошуку груп користувачів, який адаптується до розрідженості матриці користувач-предмет.

Отримав подальший розвиток метод мішаної кластеризації, який використовується для кластеризації категоріально-числових векторів профілів користувачів і, на відміну від існуючих, автоматично вибирає центри кластерів і дозволяє зменшити час пошуку груп користувачів при високій точності виділення груп.

Отримав подальший розвиток метод збільшення різноманітності рекомендованих предметів, який дозволяє врахувати оцінки подібних товарів в околі товарів активного користувача і вирішує проблему «довгого хвоста».

Удосконалено метод прогнозування рекомендацій для користувачів інтернет-магазину, який, на відміну від інших існуючих методів, використовує алгоритм пошуку асоціативних правил Apriori за допомогою адаптивної зміни підтримки асоціативних правил.

У **першому розділі** виділені основні класи електронної комерції, показано, що інтернет-магазин є одною із основних систем електронної комерції, виконана класифікація і виділені основні види інтернет-магазинів, наведені основні кроки роботи інтернет-магазину, розроблена структура роботи інтернет-магазину, показана область застосування рекомендаційних систем в структурі роботи інтернет-магазину, наведені структури Веб-сайтів для інтернет-магазину, показано зв’язок структури Веб-сайту із процесом

надання рекомендацій, наведені основні метрики ефективності роботи інтернет-магазину, виділені особливості застосування рекомендаційних систем в роботі інтернет-магазину. До таких метрик належать: кількість відвідувачів веб-сайту інтернет магазину, коефіцієнт конверсії, коефіцієнт супутніх продаж, коефіцієнт додаткових продаж. Показано, що в інтернет-магазині можна виділити три види рекомендаційних систем: вхідна рекомендаційна система, рекомендаційна система супутніх продаж, рекомендаційна система додаткових продаж. Сформульована загальна задача пошуку найкращої рекомендації: для заданої множини користувачів і заданої множини предметів рекомендаційна система для інтернет-магазину повинна рекомендувати користувачу такі предмети, які будуть відповідати його дійсним потребам.

У **другому розділі** дисертаційної роботи розроблена формальна модель задачі прогнозування рекомендацій методом колаборативної (спільної) фільтрації для інтернет-магазину. Показано, що формальна модель включає множину користувачів, множину предметів, матрицю користувач-предмет. Елементами матриці користувач-предмет є рейтингові числові оцінки, які користувачі виставляють вибраним предметам. Наведені особливості формальної моделі для прогнозування рекомендацій з урахуванням груп користувачів. Дана характеристика метрик подібності векторів в багатовимірному просторі. Виконано порівняння існуючих методів розрахунку коефіцієнтів подібності в методі зваженої суми для прогнозування рекомендацій. Проведено теоретичне і експериментальне дослідження таких метрик подібності, як косинусна відстань, коефіцієнт кореляції Пірсона, обернена евклідова відстань. Показано, що найбільшу точність дає обернена евклідова відстань. Удосконалено метод розрахунку коефіцієнтів подібності, який, на відміну від існуючих, використовує обернену евклідову відстань між векторами профілів користувачів і демографічних характеристик користувачів, показано можливість застосування цього методу для рішення задачі «холодного старту».

У **третьому розділі** дисертаційної роботи розроблено гібридний метод пошуку груп користувачів, який адаптується до розрідженості матриці користувач-предмет. Суть методу полягає в тому, що він використовує чітку кластеризацію, мішану кластеризацію і категоріальну кластеризацію. Вибір методу залежить від коефіцієнта розрідженості матриці користувач-предмет. При малій розрідженості використовується модифікований метод k-середніх, При великій розрідженості використовується двохетапний метод категоріальної і чіткої числової кластеризації. При деякому середньому значенні коефіцієнта розрідженості використовується запропонований в роботі метод мішаної кластеризації. Для сформованих груп користувачів показано застосування прогнозування рекомендацій для групи в цілому, а не для кожного окремого користувача в групі. Розроблено новий метод мішаної кластеризації, який враховує категоріальні і числові складові вектора профілю користувача і автоматично вибирає центри кластерів; показано застосування методів прогнозування рекомендацій для груп користувачів; розроблено метод прогнозування рекомендацій на основі пошуку асоціативних правил за допомогою алгоритму Apriori, який використовує алгоритм пошуку асоціативних правил за допомогою адаптивної зміни підтримки асоціативних правил. Розроблено метод прогнозування рекомендацій для супутніх продаж (cross-selling), режиму додаткових продаж (up-selling) і режиму післяпродажної роботи (e-mail marketing), розроблено метод збільшення різноманітності товарів, які пропонує інтернет-магазин і дозволяє вирішити проблему «довгого хвоста».

У **четвертому розділі** дисертаційної роботи розроблено інформаційне забезпечення для тестування моделей і методів прогнозування рекомендацій для інтернет-магазину, розроблена структура математичного забезпечення, розроблена структура програмного забезпечення, яка дозволяє вибрати метод прогнозування рекомендацій, метод пошуку груп користувачів, метод прогнозування рекомендацій в групі користувачів, метод прогнозування рекомендацій для формування додаткових продаж, метод прогнозування

рекомендацій для супутніх продаж, метод прогнозування рекомендацій для післяпродажного супроводу користувача, метод розрахунку точності прогнозування, величину поділу тестової матриці користувач-предмет на прогнозовану і тестову частини. Наведені результати експериментальних досліджень розроблених моделей, методів і алгоритмів. Експериментальні дослідження проведені на тестовому наборі даних MoviLens.

Ключові слова: рекомендаційна система, інтернет-магазин, прогнозування рекомендацій, колаборативна фільтрація, групи користувачів, мішана кластеризація, асоціативні правила.

Перелік опублікованих праць здобувача за темою дисертації:

Наукові праці, в яких опубліковані основні наукові результати дисертації:

1. Лобур М. Моделі і методи прогнозування рекомендацій для колаборативних рекомендаційних систем. / М.Лобур, М.Шварц, Ю.Стех // Вісник Національного Університету «Львівська політехніка». Інформаційні системи та мережі, Львів. – 2018. − № 901. − С. 68−75.

2. Stekh Y. Some methods for improving the accuracy of prediction recommendations / Y.Stekh, M.Lobur, M.Shvarts // Вісник Націнального Університету «Львівська політехніка». Комп’ютерні системи проетування. Теорія і практика. Львів, − 2017. − № 882. − С. 46−49.

3. Лобур М., Стех Ю., Шварц М. Метод і алгоритм прогнозування рекомендацій для спільнот користувачів / М.Лобур, Ю.Стех., М.Шварц. // Збірник наукових праць Української Академії Друкарства. Квалілогія книги. Львів, 2017. − № 1 (31). − С. 88−93.

4. Лобур М. Побудова асоціативних правил для прогнозування рекомендацій в колаборативних рекомендаційних системах / М.Лобур, Ю.Стех, М.Шварц // Збірник наукових праць Української Академії Друкарства. Квалілогія книги. Львів. – 2017. − № 2 (32). − С. 82−86.

5. Lobur M. Application of Recommender Systems in the Design of Complex Microsystem Devices / M.Lobur, M.Shvarts, Y.Stekh // International Journal of Advanced Research in Computer Engineering & Technology. − 2018. − V. 7. − № 9. − P. 709−714.

Наукові праці, які засвідчують апробацію матеріалів дисертації:

6. Shvarts M. Analysis of the Effectiveness of Similarity Measures for Recommendations Systems / M.Shvarts, M.Lobur, Y.Stekh. – In: The Experience of Design and Application of CAD Systems in Microelectronics: Proc. of the 14th International Conference, Polyana-Svalyava (Zakarpattya), 21-25 February, Lviv, 2017. − P. 275−277.

7. Shvarts M. Some Trends in Modern Recommender Systems / M.Shvarts, M.Lobur, Y.Stekh – In: Perspective technologies and methods in MEMS design: Proc. of the 13th International Conference, Polyana-Svalyava (Zakarpattya), 20-23 April. 2017, Lviv. − P. 167−169.

8. Shvarts M. Some Methods for Predicting Recommendations for MEMS Designer Communities / M.Shvarts, M.Lobur, Y.Stekh, I.Demkiv – In: Perspective technologies and methods in MEMS design: Proc. of the 14th International Conference, Polyana-Svalyava (Zakarpattya), 18-22 April, 2018, Lviv. − P. 196−199.

9. Шварц М. Моделі і методи побудови рекомендаційних систем / М.Шварц, Ю.Стех. − Проблеми та перспективи розвитку економіки і підприємництва та комп'ютерних технологій в Україні: зб. тез XIII науково-практична конференції, м.Львів, 2017, Львів. − С. 37−38.

10. Лобур М. Метод прогнозування рекомендацій з врахуванням інтересу спільноти користувачів / М.Лобур, Ю.Стех, М.Шварц. – Комп'ютерне моделювання та програмне забезпечення інформаційних систем і технологій: зб. тез третьої Всеукраїнської науково-практичної конференції, м.Рівне, 29-30 вересня, 2017, Рівне. − С. 135−137.

11. Лобур М. Використання демографічних характеристик користувачів при прогнозуванні рекомендацій / М.Лобур, Ю.Стех, М.Шварц.

− Комп’ютерне моделювання та програмне забезпечення інформаційних систем і технологій:. зб. тез третьої Всеукраїнської науково-практичної конференції, м.Рівне, 29-30 вересня, Рівне. 2017 − С. 138−139.

12. Lobur M. The method of sequential clustering for predicting recommendations / M.Lobur, M.Shvarts, Y.Stekh – In: CAD in Machinery Design-Implementation and Education Problems: Proc. of the XXV Polish-Ukrainian conference: Bielsko Biala, October 20-21, Bielsko Biala, 2017. − P. 19−20.

13. Lobur M. The Method and Algorithm for Increasing Diversity in Recommendation Systems / M.Lobur, M.Shvarts, Y.Stekh − In: CAD in Machinery Design-Implementation and Education Problems Issues: Proc. of the XXVI th International Ukrainian-Polish Scientific and Technical Conference, Lviv, 2018. − P. 110−114.

14. Kosobutskyy P. Geometric calculation of Pi using the Monte Carlo method / P.Kosobutskyy, A.Kovalchuk, M.Kuzmynykh, M.Shvarts – In: Perspective technologies and methods in MEMS design: Proc. of the 12th International Conference, Polyana-Svalyava (Zakarpattya), 20–24 April, Lviv, 2016.− P. 167−169.

ABSTRACT

Shvarts M. Hybrid models and methods of forecasting of recommendations for online store. – On the rights of the manuscript.

Dissertation for scientific degree of Candidate of Technical Sciences. Specialty 01.05.03 "Mathematical and software of computers and systems". – Lviv Polytechnic National University of Ministry of Education and Science of Ukraine, Lviv, 2019.

Contents of the dissertation. In the dissertation work the scientific task of designing and researching hybrid models and methods of forecasting of recommendations for an online store is solved, which can be used also in the operation of other subjects of e-commerce, such as electronic trading lines, Internet-shop windows, information retrieval systems, search engines in the global Internet. The main purpose of the advisory systems is to provide guidance to users when selecting items that are most relevant to their interests and preferences. Items may be goods, objects, or services.

The development and implementation of advisory systems contributed to the rapid development of information and communication technologies, namely, the Internet as a global network for storing and transmitting information, as well as the World Wide Web (WWW, Word Wide Web). At present, there are $1{,}2 \cdot 10^9$ sites on the Internet that contain $16{,}2 \cdot 10^{70}$ bytes of information. It is projected that by 2020 this figure will increase to $44 \cdot 10^{70}$ bytes. Thus, at present, the problem lies not in lack of information, but in the absence of effective mechanisms for finding information. Modern search engines (Google, Yahoo) return a much larger amount of information than the user can handle. The user may not have enough knowledge, time or experience, or both, to choose what suits his or her needs. The user explicitly or indirectly provides the system with information about their preferences. Thus, the recommendation system for the online store is presented as a system (program) that uses a certain algorithm of filtering and available information about the user's needs to recommend him a set of objects that he considers most useful to himself.

In the dissertation, the method of calculating the similarity coefficients of user profile vectors and vectors of the profile of objects, which unlike the existing, uses the demographic characteristics of the users, which allows to improve the accuracy of forecasting of recommendations and to determine the similarity coefficients for the new user and the new object.

On the basis of the concept of application in one method categorical, mixed and numeric clusterization, the method of searching for user groups was first developed, which adapts to the rarity of the user-subject matrix.

The further development of the mixed clustering method used to cluster the categorical-numeric user profile vectors and, unlike the existing ones, automatically selects cluster centers and allows you to shorten the search time for groups of users with high accuracy of group allocation.

Received further development of the method of increasing the variety of recommended items, which allows you to take into account estimates of such goods in the vicinity of the active user's products and solves the problem of "long tail".

The method of forecasting recommendations for users of the online store is improved, which, unlike other existing methods, uses the Apriori associative rules search algorithm by means of an adaptive change in the support of associative rules.

In the **First section**, the main classes of e-commerce are highlighted, the online store is shown to be one of the main e-commerce systems, the classification has been made and the main types of online stores are highlighted, the main steps of the Internet shop are outlined, the structure of the online store has been developed, the area shown the use of advisory systems in the structure of the Internet store, the structure of Web sites for the online store, shows the relationship of the structure of the Web site with the process of providing recommendations, the foundations metrics and efficiency of online store dedicated application features of recommendation systems in the online store. These metrics include: the number of visitors to the website of the online store, the conversion rate, the factor of related sales, the coefficient of additional sales. It is shown that in the online store you can distinguish three types of advisory systems: an incoming advisory system,

recommendation system of related sales, advisory system of additional sales. The general task of finding the best recommendation is formulated: for a given set of users and a given set of subjects the recommendation system for the online store should recommend to the user such items that will meet his actual needs.

In the **Second section** of the dissertation work, a formal model of the task of forecasting the recommendations is developed by the method of collaborative (common) filtering for the online store. It is shown that a formal model includes a plurality of users, a plurality of subjects, a user-object matrix. The elements of the user-subject matrix are the numerical ratings that users place on selected subjects. The features of the formal model for prediction of recommendations based on user groups are presented. A characteristic of metrics of similarity of vectors in a multidimensional space is given. Comparison of existing methods of calculation of similarity coefficients in the method of weighted sum for forecasting recommendations is made. A theoretical and experimental study of similarity metrics such as cosine distance, Pearson correlation coefficient, inverse Euclidean distance is carried out. It is shown that the reciprocal Euclidean distance gives the greatest accuracy. The method of calculating similarity coefficients, which, unlike the existing ones, uses the inverse Euclidean distance between user-user profiles and demographic characteristics of users, is improved, the possibility of using this method for solving the cold start problem is shown.

In the **Third section** of the dissertation, a hybrid method of searching for user groups is developed, which adapts to the rarity of the matrix of user-subject. The essence of the method is that it uses a clear clustering, mixed clustering, and categorical clustering. The choice of the method depends on the user-subject matrix-factor matrix. At low rarity, the modified method of k-medium is used, with a high degree of rarity, a two-stage method of categorical and clear numerical clusterization is used. With some mean value of the coefficient of rarity, the proposed method of mixed clustering is used. For formed user groups, the use of forecasting recommendations for the group as a whole is shown, not for each individual user in the group. A new method of mixed clusterization is developed

which takes into account the categorical and numerical components of the user profile vector and automatically selects cluster centers; application of methods of forecasting recommendations for groups of users is shown. The method of prediction of recommendations based on the search of associative rules is developed using the Apriori algorithm, which uses the associative rules search algorithm by means of an adaptive change in the support of associative rules. A method for forecasting recommendations for cross-selling, up-selling and post-sales mode (e-mail marketing) has been developed, a method for increasing the variety of goods offered by an online store and solving the problem of "long tail".

In the **Fourth section** of the dissertation work the information support for testing models and methods of forecasting recommendations for an online store has been developed, the structure of mathematical support has been developed, a software structure has been developed that allows you to select the method of forecasting recommendations, the method of searching for user groups, the method of forecasting recommendations in a group of users, the method forecasting recommendations for formation of additional sales, method of prediction of recommendations for related sales, method of prediction of recommendations for after sales support of the user, the method of calculating the accuracy of forecasting, the size of the division of the test matrix user-object to the predicted and test. The results of experimental studies of the developed models, methods and algorithms are presented. Experimental studies were performed on a MoviLens test data set. The results of a study of methods for calculating similarity coefficients between user profile vectors are presented. For research, the user-user weighted sum method is used. The results of a study of the accuracy of the method for forecasting recommendations for user groups depending on changes in the sparsity coefficient of the user-subject matrix are presented.

Key words: advisory system, online store, forecasting recommendations, collaborative filtering, user groups, mixed clustering, associative rules.

The list of author's publication:

Proceedings where basic scientific results where published:

1. Lobur M. Models and methods of prediction of recommendations for collaborative advisory systems / M.Lobur, M.Shvarts, Y.Stekh // Visnyk Natsionalnoho Universytetu Lvivska politekhnika. Informatsiini systemy ta merezhi, Lviv. – 2018. − № 901. − P. 68−75.

2. Stekh Y. Some methods for improving the accuracy of prediction recommendations / Y.Stekh, M.Lobur, M.Shvarts // Visnyk Natsionalnoho Universytetu Lvivska politekhnika. Kompiuterni systemy proektuvannia. Teoriia i praktyka, Lviv. – 2017. − № 882. − P. 46−49.

3. Lobur M. Method and algorithm for forecasting recommendations for user communities. / M.Lobur, Y.Stekh, M.Shvarts // Zbirnyk naukovykh prats Ukrainskoi Akademii Drukarstva. Kvalilohiia knyhy, Lviv. – 2017. − № 1 (31). − P. 88−93.

4. Lobur M. Construction of associative rules for prediction of recommendations in collaborative reference systems / M.Lobur, Y.Stekh, M.Shvarts // Zbirnyk naukovykh prats Ukrainskoi Akademii Drukarstva. Kvalilohiia knyhy, Lviv. – 2017. − № 2 (32). – P. 82−86.

5. Lobur M. Application of Recommender Systems in the Design of Complex Microsystem Devices / M.Lobur, M.Shvarts, Y. Stekh // International Journal of Advanced Research in Computer Engineering & Technology. – 2018. − V. 7. − № 9. − P. 709−714.

Proceedings that certify an approvement of thesis materials:

6. Shvarts M. Analysis of the Effectiveness of Similarity Measures for Recommendations Systems. / M.Shvarts, M.Lobur, Y.Stekh – In: The Experience of Design and Application of CAD Systems in Microelectronics: Proceedings of the 14th International Conference, Polyana-Svalyava (Zakarpattya), 21-25 February. Lviv, 2017. − P. 275−277.

7. Shvarts M. Some Trends in Modern Recommender Systems / M.Shvarts, M.Lobur, Y.Stekh – In:.Perspective technologies and methods in MEMS

design: Proceedings of the 13th International Conference, Polyana-Svalyava (Zakarpattya), 20-23 April. 2017, Lviv. − P. 167−169.

8. Shvarts M. Some Methods for Predicting Recommendations for MEMS Designer Communities / M.Shvarts, M.Lobur, Y.Stekh, I.Demkiv. – In: Perspective technologies and methods in MEMS design: Proceedings of the 14th International Conference, Polyana-Svalyava (Zakarpattya), 18-22 April, 2018, Lviv. − P. 196−199.

9. Shvarts M. Models and methods of building advisory systems // M.Shvarts, Y.Stech – In: Problems and Prospects for the Development of Economy and Entrepreneurship and Computer Technologies in Ukraine: Col. Theses of the XIII Scientific and Practical Conference, Lviv, 2017. − P. 37−38.

10. Lobur M. Method of forecasting of recommendations taking into account the interest of the user community / M.Lobur, Y.Stekh, M.Shvarts – In: Computer simulation and software of information systems and technologies: Col. Theses of the Third All-Ukrainian Scientific and Practical Conference, Rivne, 29-30 September, Rivne, 2017. − P. 135−137.

11. Lobur M. Use of demographic characteristics of users in forecasting recommendations / M.Lobur, Y.Stekh, M.Shvarts – In: Computer simulation and software of information systems and technologies: Col. Theses of the Third All-Ukrainian Scientific and Practical Conference, Rivne, 29-30 September, Rivne, 2017. − P. 138−139.

12. Lobur M. The method of sequential clustering for predicting recommendations / M.Lobur, M.Shvarts, Y.Stekh – In: CAD in Machinery Design-Implementation and Education Problems: Proceedings of the XXV Polish-Ukrainian conference: Bielsko Biala, 20-21 October, Bielsko Biala, 2017. − P. 19−20.

13. Lobur M. The Method and Algorithm for Increasing Diversity in Recommendation Systems / M.Lobur, M.Shvarts, Y.Stekh – In: CAD in Machinery Design-Implementation and Education Problems Issues: Proceedings of the XXVI th International Ukrainian-Polish Scientific and Technical Conference, Lviv, 2018. − P. 110−114.

14. Kosobutskyy P. Geometric calculation of Pi using the Monte Carlo method / P.Kosobutskyy, A.Kovalchuk, M.Kuzmynykh, M.Shvarts – In: Perspective technologies and methods in MEMS design: Proceedings of the 12th International Conference, Polyana-Svalyava (Zakarpattya), 20–24 April, Lviv, 2016. − P. 167−169.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

БД – база даних;

CBR (Case Based Reasoning) – вивід на основі прецендентів;

РС – рекомендаційна система;

КФ – колаборативна(спільна) фільтрація;

LT (Long Tail) – проблема «довгого хвоста»;

КВ – кількість відвідувачів інтернет-магазину;

БМ – Байєвська мережа;

КК – коефіцієнт конверсії;

КЛ – коефіцієнт лояльності;

САП(MAE) – середняя абсолютна похибка;

КСП – особистості коефіцієнт супутніх продаж;

КДП – аналіз коефіцієнт додаткових продаж;

КЛ – рішення коефіцієнт лояльності;

ІКФП – імовірнісна колаборативна фільтрація на основі пам'яті;

СПТ – середня пересічна точність;

НВПП – нормована відстань на основі показника продуктивності;

НСАП (NMAE) – нормована середня абсолютна похибка;

КСКП (RMSE) – коренева середньоквадратична похибка;

B2C – схема електронної комерції бізнес-споживач;

ОПХ (ROC) – отримання поточної характеристики.

ВСТУП

Актуальність теми дисертаційної роботи. Основне призначення рекомендаційних систем – надання рекомендацій користувачам при виборі предметів, які найбільше відповідають їхнім інтересам. Предметами можуть бути товари, об’єкти або послуги. Перші рекомендаційні системи були розроблені в кінці 90-х років минулого століття. Розробленню і впровадженню рекомендаційних систем сприяв стрімкий розвиток інформаційно-комунікаційних технологій, а саме, Інтернет, як всесвітньої мережі для зберігання і передачі інформації, а також Всесвітньої павутини (WWW, Word Wide Web). На даний час в Інтернеті зберігається $1{,}2 \cdot 10^{9}$ сайтів, які містять $16{,}2 \cdot 10^{70}$ байтів інформації. Прогнозується, що до 2020 року цей показник зросте до $44 \cdot 10^{70}$ байтів. Таким чином на даний час проблема полягає не у відсутності інформації, а у відсутності ефективних механізмів пошуку інформації. Сучасні пошукові системи (Google, Yahoo) повертають значно більший об’єм інформації, ніж користувач здатний її обробити. Користувачеві може не вистачати знань або досвіду, або того і іншого, щоб вибрати те, що відповідає його потребам. Користувач явно або побічно надає системі інформацію про свої уподобання. Таким чином, рекомендаційна система представляється у вигляді системи (програми), що використовує певний алгоритм фільтрації та наявну інформацію про потреби користувача, щоб рекомендувати йому набір об'єктів, які він вважає найбільш корисними для себе. Рекомендаційні системи використовуються в середовищі електронної комерції, пошукових системах, системах електронної освіти. Рекомендаційні системи класифікують за способом відбору необхідного користувачеві матеріалу. В основному застосовується два базові підходи: коллаборативна фільтрація і контентна фільтрація. Також існує гібридна фільтрація, яка поєднує в собі як колаборативну, так і контентну фільтрацію. У рекомендаційних системах, які використовують контентну фільтрацію (фільтрація по змісту), користувачі не залежать від інших користувачів

системи. Для формування рекомендацій системі необхідний профіль користувача з інформацією про його інтереси. У профілі в певній формі зберігається інформація про об' єкти, до яких користувач вже виявив інтерес. Система також містить в себе інформацію про всі предмети, які вона може рекомендувати. Така система використовує опис об'єктів в профілі користувача, знаходить схожі об'єкти в своїй базі даних, а потім рекомендує їх йому. Застосування фільтрації такого роду дуже доречне, коли користувач має чітко визначені, конкретні інтереси і шукає схожі рекомендації. Перевага контентної фільтрації полягає в тому, що для початку надання рекомендацій не потрібна велика кількість зареєстрованих користувачів, тобто рекомендації не залежать від інших користувачів системи. Основним обмеженням даного методу є неможливість системи з таким видом фільтрації рекомендувати нові об'єкти, які не відповідають інтересам користувача. Рекомендаційні системи в найбільшій мірі застосовуються в інтернет-магазинах. 03.09.2015 року Верховна Рада України прийняла Закон України « Про електронну комерцію » [15]. В цьому Законі надано наступне визначення інтернет- магазину – це засіб представлення або реалізації товару, роботи чи послуги шляхом вчинення електронного правочину. Важливий внесок в розроблення структури і функцій інтернет-магазину здійснили такі вчені, як Плескач В.Л., Затонацька Е.Г. – розробили тактичні прийоми електронної комерції, дали характеристику поняття електронного магазину, розробили типову структуру інтернет-магазину[16,17], Шердані А. – розробив метод комплексного аналізу і порівняння економічної ефективності інтернет-магазинів [18], Шалева О.І. – дослідила основні категорії та інструментарій електронної комерції, технологію роботи інтернет-магазину [19], Тардаскіна Т.М., Стрельчук Є.М., Терешко Ю.В. – дослідили системи електронної комерції в споживчому секторі, виділили інтернет-магазин як один із основних елементів системи електронної комерції в споживчому секторі [20], Huang S.I. – дослідив застосування рекомендаційних систе в електронній комерції [21], Linden G., Smith J., York J. – розглянули застосування методу колаборативної фільтрації в

інтернет-магазині Amazon.com [22], Resnick P. – розробив одну із перших рекомендаційних систем GroupLens [23].

Основними задачами рекомендаційних систем в області електронної комерції, в тому числі і при функціонуванні інтернет-магазину, є наступні: збільшення достовірності прогнозування рекомендацій, вирішення проблеми «холодного старту», підвищення різноманітності рекомендованих предметів.

Для інтернет-магазину додатковими задачами є збільшенням конверсії, збільшенням лояльності користувачів, вирішення задачі супутніх продаж (cross-selling), вирішення задачі додаткових продаж (up-selling), вирішення задачі післяпродажного супроводу користувачів (e-mail маркетинг), вирішення задачі «довгого хвоста» (long tail). Окрім класичних задач, які перечислені вище, актуальною задачею підвищення ефективності роботи інтернет-магазину є задача пошуку груп користувачів з подібними уподобаннями і демографічними характеристиками. Розбиття користувачів на такі групи дозволяє підвищити точність прогнозування рекомендацій, вирішити задачу «холодного старту», а також полегшує вирішення задач супутніх продаж, додаткових продаж і після продажного супроводу.

У рекомендаційних системах для інтернет-магазинів точність прогнозування рекомендацій залежить від точності розрахунку коефіцієнтів подібності векторів характеристик користувачів і векторів характеристик предметів. Вектори характеристик користувачів і вектори характеристи предметів включають числові оцінки корисності предметів для користувачів. Актуальною є задача розроблення і дослідження подібності векторів характеристик користувачів і векторів подібності характеристик предметів при врахуванні демографічних харвавктеристик користувачів і якісних характеристик предметів.

Для ефективного функціонування інтернет-магазинів, особливо великих торгових центрів (супермаркетів, гіпермаркетів) актуальною є задача пошуку груп користувачів і груп предметів з подібними характеристиками. Це дозволяє замінити задачу прогнозування рекомендацій для одного

користувача на задачу прогнозування рекомендацій для груп користувачів і задачу прогнозування рекомендацій для одного предмета на задачу прогнозування рекомендацій для груп предметів. До задач для яких доцільно використовувати прогнозування для груп предметів, відноситься задача прогнозування супутніх продаж (cross-selling), задача прогнозування додаткових продаж (up-selling), задача після продажного супроводу користувача (e-mail маркетинг), задача прогнозування телевізійних програм, музичних програм для певних груп користувачів. Задачі такого типу також виникають при проектуванні складних систем різного фунціонального призначення колективами проектувальників. Тому актуальною є задача виділення груп користувачів з близькими інтересами і вибір методу прогнозування рекомендацій для кожної групи.

Для вирішення задачі виділення груп користувачів і груп предметів з подібними характеристиками у більшості випадків використовують методи кластеризації. Кластеризація дозволяє отримати декілька матриць користувач-предмет значно меншої розмірності, ніж початкова матриця і меншої розрідженості. Однак ефективність розбиття на кластери в значній мірі залежить від розрідженості матриці користувач-предмет. При малій розрідженості можна ефективно використовувати такі методи чіткої чисельної кластеризації, як k-середніх і його модифікації. При великій розрідженості методи чіткої чисельної кластеризації дають незадовільний результат. При великій розрідженості матриці користувач-предмет доцільно використовувати двоетапну кластеризацію. На першому етапі виконується категоріальна кластеризація матриці користувач-предмет на основі категоріальних векторів профілів користувачів. На другому етапі виконується чітка числова кластеризація кожної із отриманих матриць. При середньому значенні величини розрідженості матриці користувач-предмет використовується мішана кластеризація. При цьому чисельний вектор профілю користувача доповнюється категоріальними демографічними характеристиками користувача. Для кластеризації категоріально-числових векторів профілів

користувачів розроблений метод мішаної кластеризації, який автоматично вибирає центри кластерів і дозволяє зменшити час пошуку груп користувачів при високій точності виділення кластерів. Тому актуальною є задача розроблення гібридних методів прогнозування рекомендацій, які адаптуються до розрідженості матриці користувач-предмет.

Важливою задачею в сучасних рекомендаційних системах є підвищення різноманітності прогнозованих предметів. Суть цієї задачі полягає в тому, що велика розрідженість матриці користувач-предмет призводить до того, що користувачу рекомендаційна система рекомендує предмети з близькими характеричстиками. Це призводить до того, що ряд корисних для користувача предметів залишається поза його увагою.

Близькою до попередньої задачі є вирішення задачі «довгого хвоста» (long tail). Суть задачі «довгого хвоста» полягає в тому, що користувачам пропонують переважно предмети високої популярності. Кількість таких предметів не перевищує 20% від загальної кількості предметів, які може запропонувати інтернет-магазин. Тому актуальним є завдання прогнозування користувачу предметів, які належать до «довгого хвоста» і мають властивості, які подібні до вибраних предметів з високою популярністю.

У **вступі** обґрунтовано актуальність розроблення гібридних моделей і методів прогнозування рекомендацій для інтернет-магазинів як суб'єктів електронної комерції, сформульовано мету і завдання роботи, основними з яких є розроблення формальних моделей і методів пошуку груп предметів і користувачів з подібними характеристиками, вступ в процес прогнозування рекомендацій на основі демографічних характеристик користувачів і якісних характеристик предметів, збільшення різноманітності прогнозованих предметів.

У **першому розділі** виділені основні класи електронної комерції, показано, що інтернет-магазин є однією із основних систем електронної комерції, виконана класифікація і виділені основні види інтернет-магазинів, наведені основні кроки роботи інтернет-магазину, розроблена структура

роботи інтернет-магазину, показана область застосування рекомендаційних систем в структурі роботи інтернет-магазину, наведені структури Веб-сайтів для інтернет-магазину, показано зв’язок структури Веб-сайту із процесом надання рекомендацій, наведені основні метрики ефективності роботи інтернет-магазину, виділені особливості застосування рекомендаційних систем в роботі інтернет-магазину.

У **другому розділі** дисертаційної роботи розроблена формальна модель задачі прогнозування рекомендацій методом колаборативної фільтрації для інтернет-магазину, приведено особливості формальної моделі для прогнозування рекомендацій з урахуванням груп користувачів, виконано порівняння існуючих методів розрахунку коефіцієнтів подібності в методі зваженої суми, удосконалено метод розрахунку коефіцієнтів подібності, який, на відміну від існуючих, використовує обернену евклідову відстань між векторами профілів користувачів і демографічні характеристики користувачів, показано можливість застосування цього методу для рішення задачі «холодного старту».

У **третьому розділі** дисертаційної роботи розроблено гібридний метод пошуку груп користувачів, який адаптується до розрідженості матриці користувач-предмет; розроблено новий метод мішаної кластеризації, який враховує категоріальні і числові складові вектора профілю користувача і автоматично вибирає центри кластерів; показано застосування методів прогнозування рекомендацій для груп користувачів; розроблено метод прогнозування рекомендацій на основі пошуку асоціативних правил, який використовує алгоритм пошуку асоціативних правил за допомогою адаптивної зміни підтримки асоціативних правил; розроблено метод прогнозування рекомендацій для супутніх продаж (cross-selling), режиму додаткових продаж (up-selling) і режиму післяпродажної роботи (e-mail marketing), розроблено метод збільшення різноманітності товарів, які пропонує інтернет-магазин і дозволяє вирішити проблему «довгого хвоста» а саме: це 80% предметів, які

не користуються великою популярністю, але можуть представляти інтерес для користувачів.

У **четвертому розділі** дисертаційної роботи розроблено інформаційне забезпечення для тестування моделей і методів прогнозування рекомендацій для інтернет-магазину, розроблена структура математичного забезпечення, розроблена структура програмного забезпечення, яка дозволяє вибрати метод прогнозування рекомендацій, метод пошуку груп користувачів, метод прогнозування рекомендацій в групі користувачів, метод прогнозування рекомендацій для формування додаткових продаж, метод прогнозування рекомендацій для супутніх продаж, метод прогнозування рекомендацій для післяпродажного супроводу користувача, метод розрахунку точності прогнозування, величину поділу тестової матриці користувач-предмет на прогнозовану і тестову.

Зв'язок роботи з науковими програмами, планами, темами

Дисертаційна робота виконана на кафедрі «Системи автоматизованого проектування» Національного університету «Львівська політехніка» та безпосередньо пов'язана з планами наукових досліджень в рамках госпдоговірних тем та міжнародних наукових грантів, і відповідає науковому напряму кафедри «Системи автоматизованого проектування» Національного університету «Львівська політехніка» - Автоматизація проектування та моделювання систем «розумного будинку» та виконана в межах госпдоговірних тем та міжнародних наукових грантів, а саме:

- міжнародний науково-дослідний проект − TEMPUS-JPCR – «Розробка програми для нової спеціальності: "Магістр з інженерії проектування мікросистем"», термін виконання проекту 15.10.2012 - 14.10.2016, реєстраційний номер: № 530785-«TEMPUS-1-2012-1-PL- TEMPUS-JPCR»;
- грант молодих учених − ДБ/Нанокомпозит «Моделювання і створення нового класу кристалічних нанокомпозитів із контрольованою кристалізацією та їх дослідження в оптичному та субтерагерцовому діапазонах

хвиль», термін виконання проекту 01.01.2016 - 31.12.2018, номер державної реєстрації № 0116U004412.

Мета та завдання дослідження

Метою дисертаційної роботи є вироблення релевантних прогнозів рекомендацій для кінцевих користувачів інтернет-магазину за рахунок розроблення і удосконалення моделей та методів прогнозування рекомендацій, спрямованих на підвищення точності, достовірності (релевантності) і швидкості прогнозування рекомендацій.

Для досягнення цієї мети були поставлені такі завдання:

- аналіз сучасного стану моделей і методів прогнозування рекомендацій для задач електронної комерції та інтернет-магазинів як суб’єктів електронної комерції;
- удосконалення методів розрахунку коефіцієнтів подібності векторів профілів користувачів і предметів для методу зваженої суми шляхом використання демографічних характеристик користувачів, якісних характеристик предметів і оберненої евклідової відстані між векторами профілів;
- розроблення методу пошуку груп користувачів, який адаптується до коефіцієнту розрідженості матриці користувач-предмет і використовує методи категоріальної, мішаної і числової кластеризації;
- розроблення методу мішаної кластеризації, який використовується для кластеризації категоріально-числових векторів профілів користувачів і автоматично вибирає центри кластерів;
- розроблення методу збільшення різноманітності рекомендованих предметів;
- розроблення методу прогнозування рекомендацій для користувачів інтернет-магазину на основі концепції асоціативних правил, який використовує алгоритм пошуку асоціативних правил за допомогою адаптивної зміни підтримки асоціативних правил;

- дослідження розроблених моделей і методів на тестовому наборі даних Movilens.

Об'єкт дослідження − процес вироблення рекомендацій інтернет-магазином для користувача.

Предмет дослідження **–** моделі, методи та алгоритми прогнозування рекомендацій для користувачів інтернет-магазину.

Методи дослідження – поставлені в дисертаційній роботі завдання розв'язувалися за допомогою: методів лінійної алгебри, методів кластерного аналізу, теорії реляційних баз даних, методів пошуку асоціативних правил, математичної статистики.

Наукова новизна отриманих результатів полягає в тому, що:

1. На основі концепції застосування в одному методі категоріальної, мішаної і числової кластеризації вперше розроблено метод пошуку груп користувачів, який адаптується до розрідженості матриці користувач-предмет.
2. Отримав подальший розвиток метод розрахунку коефіцієнтів подібності векторів профілів користувачів і векторів профілів предметів, який, на відміну від існуючих, використовує демографічні характеристики користувачів, що дозволяє підвищити точність прогнозування рекомендацій і визначати коефіцієнти подібності для нового користувача і нового предмета.
3. Отримав подальший розвиток метод мішаної кластеризації, який використовується для кластеризації категоріально-числових векторів профілів користувачів і, на відміну від існуючих, автоматично вибирає центри кластерів і дозволяє зменшити час пошуку груп користувачів при високій точності виділення груп.
4. Отримав подальший розвиток метод збільшення різноманітності рекомендованих предметів, який дозволяє врахувати оцінки подібних предметів в околі предметів активного користувача і вирішує проблему «довгого хвоста».

5. Удосконалено метод прогнозування рекомендацій для користувачів інтернет-магазину, який на відміну від існуючих методів, використовує алгоритм пошуку асоціативних правил за допомогою адаптивної зміни підтримки асоціативних правил.

Практичне значення отриманих результатів полягає в наступному: розроблені моделі і методи можуть бути застосовані в електронній комерції при створені інтернет-магазинів, інтернет-аукціонів, веб-порталів, інформаційно-пошукових системах.

Результати дисертаційних досліджень використано у навчальному процесі Національного університету «Львівська політехніка», кафедри систем автоматизованого проектування у дисциплінах: «Інноваційні інформаційні технології» для підготовки магістрів за спеціальністю «Інформаційні технології проектування» і «Методи і системи штучного інтелекту» для підготовки бакалаврів за спеціальністю «Комп'ютерні науки».

Особистий внесок здобувача

Усі наукові результати дисертаційної роботи отримані здобувачем особисто. У друкованих працях, написаних у співавторстві, здобувачеві належать:

в [1] − розроблено метод пошуку груп користувачів, який адаптується до розрідженості матриці користувач-предмет і використовує категоріальну, мішану і числову кластеризацію;

в [2] − вдосконалено метод обчислення коефіцієнтів подібності векторів характеристик користувачів і предметів за рахунок введення демографічних і якісних характеристики користувачів і предметів;

в [3] − розроблено метод надання рекомендацій для груп користувачів на основі кластеризації векторів профілів користувачів;

в [4] − розроблено метод, який надає рекомендації новому користувачу, пропонує користувачу супутні предмети і використовує асоціативні правила;

в [5] − розроблено метод прогнозування рекомендацій при проектуванні складних мікросистемних пристроїв;

в [6] − проведено аналіз методів розрахунку коефіцієнтів подібності характеристик користувачів і предметів. Показано недоліки косинусної міри подібності, коефіцієнта кореляції Пірсона. Експериментально доведено ефективність оберненої евклідової відстані як міри подібності векторів;
в [7] − проведено дослідження стану моделей, методів, засобів і алгоритмів побудови рекомендаційних систем;
в [8] − розроблено метод прогнозування рекомендацій при проектуванні мікроелектромеханічних систем;
в [9] − досліджені основні концепції побудови і застосування рекомендаційних систем;
в [10] − досліджена модель прогнозування рекомендацій для груп користувачів на основі кластеризації векторів профілів користувачів, які входять в матрицю користувач-предмет;
в [11] − удосконалено метод розрахунку коефіцієнтів подібності користувачів за рахунок демографічних характеристик;
в [12] − розроблено метод двоетапної категоріальної і числової кластеризації для виділення груп користувачів;
в [13] − розроблено метод збільшення різноманітності рекомендованих предметів;
в [14] − проведено аналіз геометричного розрахунку Пі за допомогою існуючого методу «Монте Карло».

Апробація результатів дисертації

Основні результати дисертаційної роботи доповідалися і обговорювалися на семінарах та конференціях:

- Міжнародній науково-технічній конференції «Перспективні технології і методи проектування МЕМС» (Львів-Поляна, 2016, 2017, 2018);
- Міжнародній науково-технічній конференції «Досвід розробки та застосування САПР в мікроелектроніці» (Львів-Поляна, 2017);
- Міжнародній науково-технічній конференції «САПР у проектуванні машин. Питання впровадження і навчання» (Львів-Поляна, 2017, 2018);

- Науково-практичній конференції «Проблеми та перспективи розвитку економіки і підприємництва та комп’ютерних технологій в Україні» (Львів, 2017);
- Всеукраїнській науково-практичній конференції «Комп’ютерне моделювання та програмне забезпечення інформаційних систем і технологій» (Рівне, 2017);
- наукових семінарах кафедри «Системи автоматизованого проектування» Національного університету «Львівська політехніка» (2015−2018).

Публікації

Основні результати дисертаційної роботи висвітлено в 14 наукових публікаціях, з яких: 1 наукова стаття в науковому періодичному виданні іншої держави, яке входить в міжнародні науково метричні бази, 4 статті опубліковано у фахових виданнях України, які входять в перелік фахових видань МОН України, 9 тез доповідей в матеріалах науково-технічних конференцій, 6 з яких міжнародні.

Структура та обсяг дисертації

Дисертаційна робота складається із вступу, чотирьох розділів, висновків, списку літературних джерел , та додатків. Загальний обсяг дисертації складає 152 сторінки, з них 135 сторінок основного тексту, містить 41 рисунок та 10 таблиць. Список літератури містить 113 найменувань.

РОЗДІЛ 1. АНАЛІЗ СУЧАСНОГО СТАНУ МОДЕЛЕЙ І МЕТОДІВ ПРОГНОЗУВАННЯ РЕКОМЕНДАЦІЙ В РЕКОМЕНДАЦІЙНИХ СИСТЕМ ІНТЕРНЕТ-МАГАЗИНІВ

1.1. Електронний бізнес, електронна комерція, основні класи електронної комерції

На даний час Інтернет є найбільшим у світі об’єднанням локальних мереж і забезпечує швидкий і надійний обмін інформацією між користувачами. Швидкість, надійність і доступність Інтернету спричинила появу нового напрямку в інформаційно-комунікаційних технологіях початку XXI століття і світовій економіці електронного бізнесу в цілому.

Бізнес – діяльність з виробництва і реалізації товарів і послуг, яка здійснюється в умовах конкуренції на ринку й метою якої є отримання прибутку. Сутність бізнесу полягає у поєднанні інтелектуальних, матеріальних, фінансових, трудових, інформаційних ресурсів з метою виробництва і продажу товарів або послуг громадянам, компаніям, організаціям [16,17,24,25].

Концепція е-бізнесу виникла у США у 80-х роках XX ст. і стала результатом розвитку ідеї глобальної інформаційної економіки і застосування інформаційних технологій (IT) в компаніях [16,17,19]. У ролі суб’єктів електронного бізнесу можуть виступати компанія, корпорація, держава і тому подібне. Інформаційний сектор економіки є основою для зазначеної трансформації традиційних форм господарювання в економічну систему постіндустріального типу.

Електронний бізнес – ділова активність, що використовує можливості глобальних інформаційних мереж для перетворення внутрішніх і зовнішніх зв'язків компанії з метою створення прибутку.

Електронна комерція – це один зі способів здійснення електронного бізнесу. Це різновид бізнес-активності, в якій комерційна взаємодія суб'єктів

бізнесу з купівлі-продажу товарів і послуг (як матеріальних, так і інформаційних) здійснюється з допомогою Internet або будь-якої іншої інформаційної мережі. З 03.09.2015 року в Україні діє Закон « Про електронну комерцію »[15].

1.2. Системи електронної комерції, які розраховані на взаємодію бізнесу і користувача (споживчий сектор (B2C))

До таких систем електронної комерції належать [26-31]:

- Електронна крамниця;
- Електронний довідник-каталог;
- Електронний он-лайновий аукціон;
- Електронний торговельний центр;
- Віртуальне співтовариство;
- Віртуальний центр розробки;
- Інформаційний брокер;
- Провайдер бізнес-операцій;
- Іінтегратор бізнес-операцій тощо

1.3. Структурна схема функціонування інтернет-магазину

Типова схема взаємодії покупця з інтернет-магазином здійснюється наступним чином (рис.1.1). Покупець за допомогою Application заходить на сайт Інтернет-магазину і виконує наступні дії:

1. Формування кошика покупця шляхом перегляду товарного каталогу та вибору товарів;
2. Реєстрація покупця;
3. Покупець вибирає форму доставки та оплати товару;
4. Підтвердження замовлення;
5. Оплата товару;

6. Доставка придбаного товару покупцеві.

На рис.1.1 наведена схема функціонування інтернет-магазину без рекомендаційної системи.

Рис.1.1 – Структура функціонування інтернет-магазину без рекомендаційних систем

Типова схема взаємодії покупця з Інтернет-магазином при наявності рекомендаційних систем наведена на рис. 1.2.

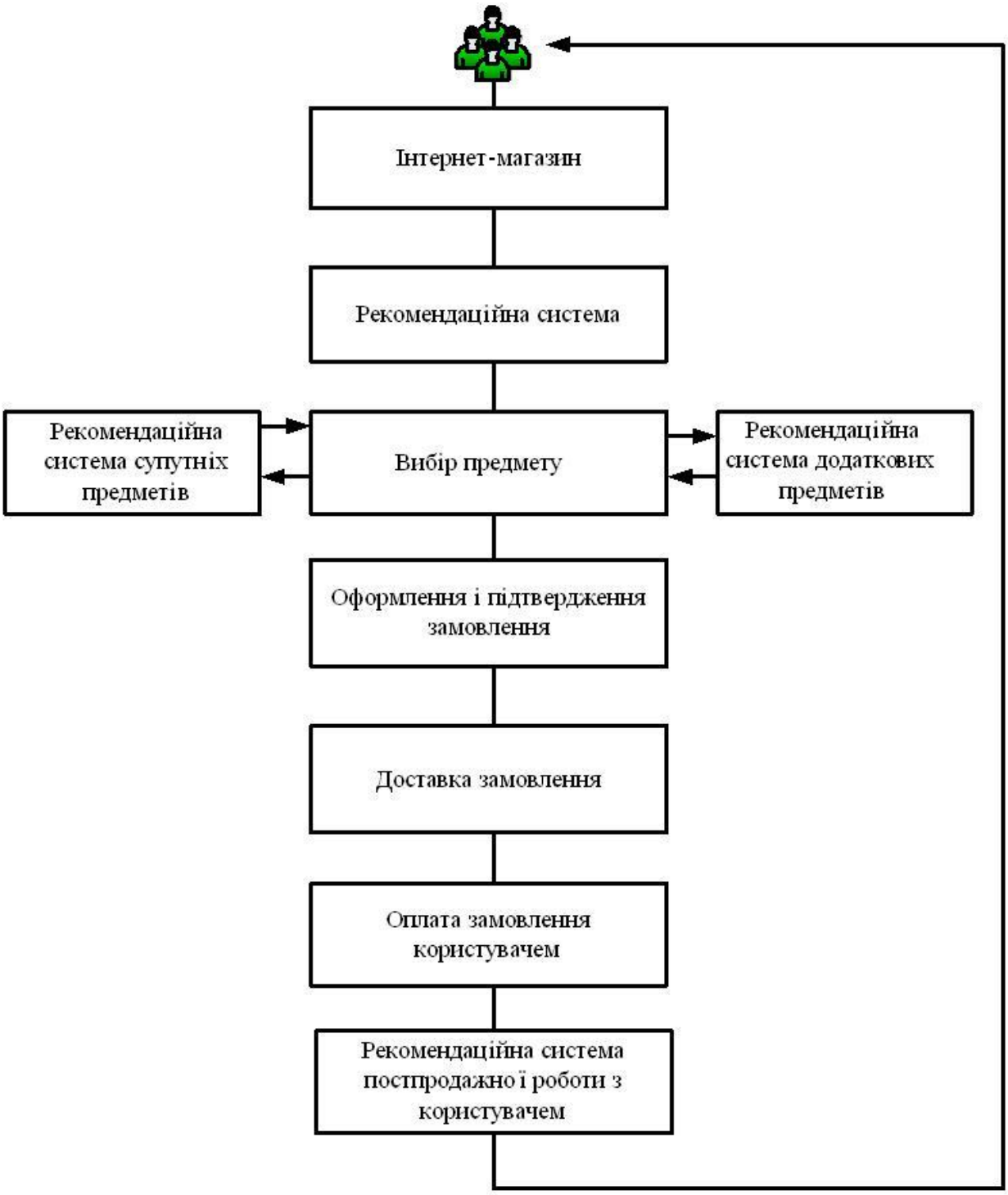


Рис. 1.2 – Структура функціонування інтернет-магазину при наявності рекомендаційних систем

Рекомендаційна система призначена для рекомендації користувачу предметів, які найбільше відповідають його уподобанням. Рекомендації будуються на основі профілів користувачів, які містять інформацію про оцінки раніше вибраних користувачем предметів, інформацію про персональні характеристики користувача, а також на основі профілів предметів, які містять інформацію про оцінки користувачами даного предмету, інформацію про контентні характеристики предметів.

Рекомендаційна система супутніх предметів призначена для рекомендації користувачу предметів, які вибирали інші користувачі, але інших категорій. Такі рекомендації будуються на основі асоціативних правил виду «якщо вибраний j-тий предмет, то разом із ним вибирають предмети «j+1, j+2,…j+k».

Рекомендаційна система додаткових предметів призначена для поповнення споживчого кошика предметами тієї ж категорії, що і вибраний предмет, але вищої якості і відповідно вищої ціни.

1.4. Структури Веб-сайтів для інтернет-магазину

Веб-сайт інтернет-магазину є однією з основних структурних компонент інтернет-магазину. Від структури і інформаційного наповнення Веб-сайту в значній мірі залежить успіх інтернет-магазину [27−34].

Структура сайту - логічні блоки, збудовані в певній послідовності. Від системи розміщення сторінок розділів, карток товарів, інформаційних блоків залежить зручність переміщення користувача по сайту, швидкість пошуку і ймовірність знаходження потрібного, а отже, продажу. Адже якщо клієнт не знаходить хліб в хлібному відділі, він взагалі йде з супермаркету.

Правильно створити структуру для інтернет-магазину - означає зберегти потенційних клієнтів, поліпшити показники пошукового просування і отримати довіру пошукових систем.

1.4.1. Типи структур сайтів

Основні види структур - лінійна, лінійна з розгалуженнями, блокова структура, деревоподібна структура і тегова структура [32−34]. Ми розглянемо тільки останні дві, оскільки вони найкраще підходять для створення інтернет-магазину. Тоді як варіанти лінійної і блокової структур застосовуються для сайтів-презентацій, портфоліо, онлайн-книг, продажу одного продукту або послуги.

1.4.2. Деревоподібна структура

Структура має на увазі вкладеність однієї категорії в іншу, для кожної послуги або товару формується окрема гілка (рис.1.3). Це стосується брендів, категорій, видів продукції.

Більшість сайтів має саме таку структуру, на них ми бачимо звичні розділи, підрозділи і окремі товари. Подібний формат найбільш зручний для сприйняття:

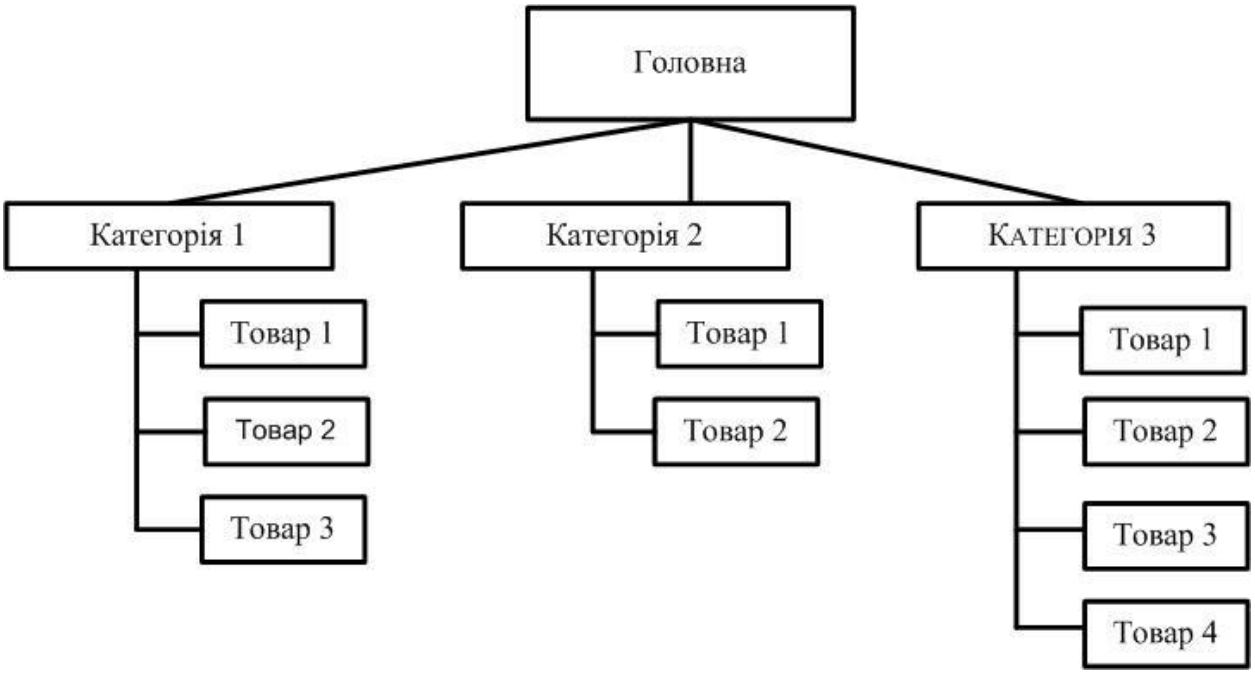


Рис. 1.3 – Деревоподібна структура Веб-сайту інтернет-магазину

1.4.3. Структура, яка складається із тегів

Структура, при якій створюються окремі сторінки тегів на основі певних параметрів: ціна, акції, особливості використання, властивості, специфічні

характеристики і так далі. На прикладі категорії «мобільні телефони» тегами можуть бути: «телефони з 2 sim», «з великою ємністю АКБ», «бюджетні».

Перевага структури, яка складається із тегів - збільшення кількості сторінок, які можуть залучити додатковий трафік. Такий результат досягається за рахунок розширення семантичного ядра: використання більшої кількості низькочастотних запитів.

Структура, яка складається із тегів може бути накладена на деревоподібну. Схематично це виглядає так (рис.1.4):

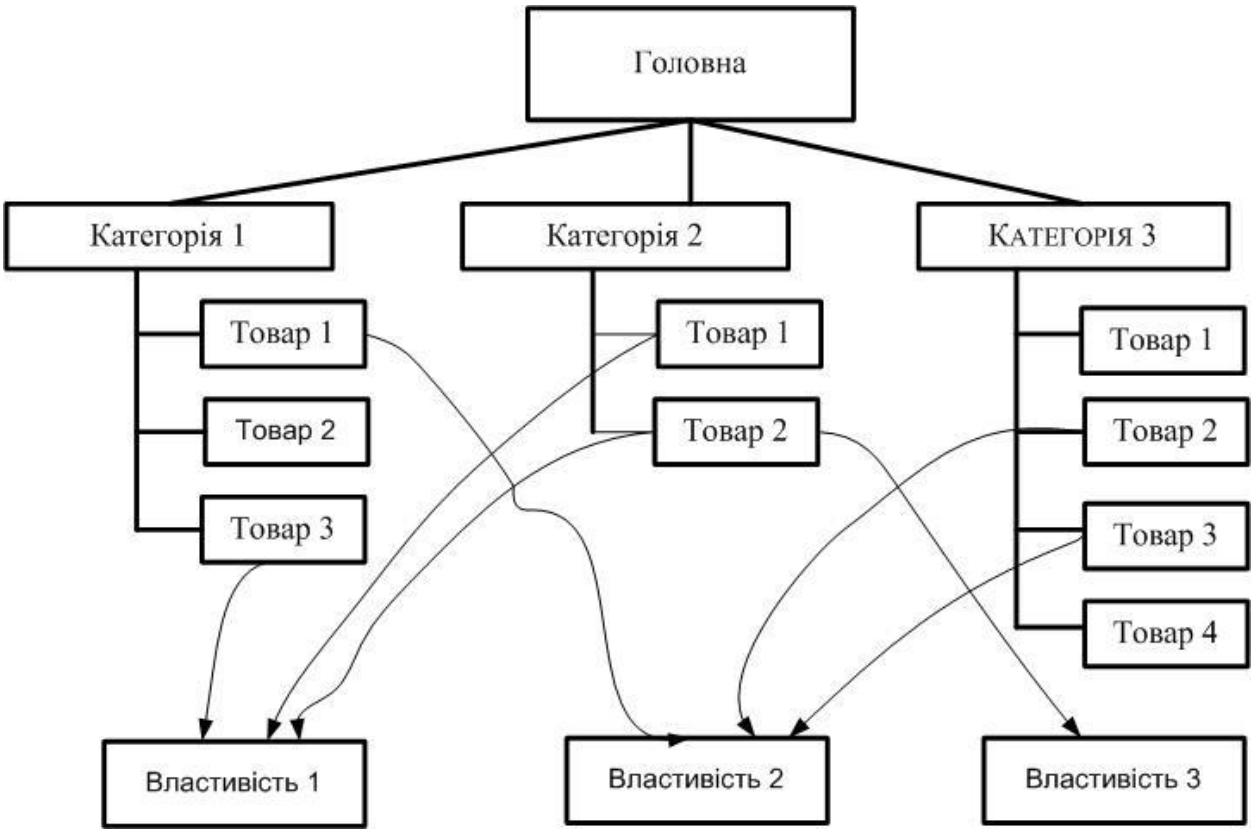


Рис.1.4 – Деревоподібна структура Веб-сайту інтернет-магазину з накладеними тегами

1.4.4. Вимоги до оптимальної структури сайту інтернет-магазину

Правильна структура інтернет-магазину складається таким чином, щоб вона задовольняла користувача в зручності використання і пошукових роботів для ефективного і швидкого індексування. Тому необхідно дотримуйтися основних вимог [33,34]:

- логічність;

- невеликий рівень вкладеності;
- зрозумілі назви;
- розміщення одного товару в одній категорії, а не в декількох.

Структура сайту в значній мірі впливає на ефективність роботи рекомендаційних систем. Для вхідної рекомендаційної системи і рекомендаційної системи супутніх товарів доцільною є деревоподібна структура. Для рекомендаційної системи додаткових товарів більш доцільною є структура, яка складається з тегів.

1.5. Основні метрики ефективності інтернет-магазинів

До основних метрик ефективності інтернет-магазину, які залежать від ефективної роботи рекомендаційних систем можна віднести [18]:

- Кількість відвідувачів інтернет-магазину;
- Коефіцієнт конверсії;
- Коефіцієнт лояльності;
- Коефіцієнт супутніх продаж;
- Коефіцієнт додаткових продаж.

1.5.1. Кількість відвідувачів інтернет-магазину

Це досить суб’єктивна метрика. Викликано це тим, що користувач може лише увійти на сайт інтернет-магазину і одразу вийти з нього не вчинивши жодної дії. Однак цей показник показує кількість можливих потенційних користувачів інтернет-магазину.

1.5.2. Коефіцієнт конверсії

Поняття конверсії має досить широке застосування в різних галузях науки і пракики: психології, лінгвістиці, мікробіології. В інтернет-маркетингу під конверсією розуміють частку візитів користувачів на веб-сайт, в ході яких відвідувачі здійснили покупку предмету в споживчий кошик, відвідування певної сторінки сайту, відправку заявки через форму зворотнього зв’язку,

покупку предмета. Для кожного виду цільової дії можна розрахувати свій коефіцієнт конверсії. Для кожного виду конверсії можна розрахувати свій коефіцієнт конверсії за наступною формулою:

$$C_{conv} = \frac{C_{action}}{FC_{action}} * 100\%, \quad (1.1)$$

де C_{conv} – коефіцієнт конверсії;

C_{action} – кількість дій певного типу;

FC_{action} – загальна кількість дій.

Найбільш ефективною метрикою є коефіцієнт конверсії користувача в споживача. Він показує яка доля відвідувачів сайту інтернет-магазину купила предмети. На величину коефіцієнта конверсії впливає ефективність роботи вхідної рекомендаційної системи.

1.5.3. Коефіцієнт лояльності

Лояльність користувача – це повторне звернення користувача до інтернет-магазину для придбання предметів. Коефіцієнт лояльності можна розрахувати за наступною формулою:

$$C_{loyal} = \frac{C_{repeat}}{FC_{custom}} * 100\% \ , \quad (1.2)$$

де C_{repeat} – кількість користувачів, які здійснили повторне звернення до інтернет-магазину;

FC_{custom} – загальна кількість користувачів, які здійснили звернення до інтернет-магазину. На величину коефіцієнта лояльності впливає ефективність роботи вхідної рекомендаційної системи.

1.5.4. Коефіцієнт супутніх продаж

Супутні продажі - це предмети, які придбав користувач із множини предметів інших категорій, рекомендованих рекомендаційною системою супутніх предметів. Коефіцієнт супутніх продаж можна розрахувати за наступною формулою:

$$FC_{comp} = \frac{C_{comp}}{FC_{buy}}, \quad (1.3)$$

де C_{comp} – кількість предметів, які придбав користувач із множини предметів, рекомендованих рекомендаційною системою супутніх предметів;

FC_{buy} – загальна кількість предметів, які придбав користувач.

На величину коефіцієнта супутніх товарів впливає ефективна робота рекомендаційної системи супутніх товарів.

1.5.5. Коефіцієнт додаткових продаж

Додаткові продажі – це предмети, які придбав користувач із множини предметів тих самих категорій, що і основні придбані предмети і які рекомендувала рекомендаційна система додаткових продаж. Коефіцієнт додаткових продаж можна обчислити за наступною формулою:

$$FC_{addit} = \frac{C_{addit}}{FC_{buy}}, \quad (1.4)$$

де C_{addit} – кількість предметів, які придбав користувач із множини предметів рекомендованих рекомендаційною системою додаткових предметів;

FC_{buy} – загальна кількість предметів, які придбав користувач.

На величину коефіцієнта додаткових продаж впливає впливає ефективна робота рекомендаційної системи додаткових продаж.

1.6. Методи рішення задач прогнозування рейтингів

Теорія і практика побудови рекомендаційних систем ґрунтується на теорії інформаційного пошуку [36], теорії пізнання [22], теорії маркетингу [37], теорії прогнозування [38], теорії управління [39]. Рекомендаційні системи призначені для прогнозування невідомих рейтингових оцінок предметів на основі подібних профілів користувачів [40]. Узагальнену модель функціонування рекомендаційної системи можна представити наступним

чином [41]. Нехай $\mathbf{U}$ – множина користувачів, $\mathbf{I}$ – множина предметів, які можуть бути рекомендовані, $\mathbf{F}$ – функція, яка відображає корисність предмета i для користувача u . Тоді для користувача $u \in \mathbf{U}$ система рекомендує такий предмет $i' \in \mathbf{I}$, котрий максимізує значення функції $\mathbf{F}$:

$$i' = \arg\max_{c \in \mathbf{C}\ i \in \mathbf{I}} \mathbf{F}(c, i). \tag{1.5}$$

Кожнен вектор множини $\mathbf{U}$ є профілем користувача. Профіль користувача містить різноманітні характеристики користувача, наприклад: стать, освіта, вік, професія, рід занять, множину числових рейтингових оцінок, які користувач виставив вже вибраним предметам.

Кожнен вектор множини $\mathbf{I}$ є профілем предмета і містить характеристики предмета. Наприклад для книги це може бути назва жанру, ім’я автора, рік видання, назва книги, а також вектор може складатися з числових рейтингових оцінок, які користувачі виставили предмету.

Процес прогнозування рекомендацій в рекомендаційній системі може складається із двох підзадач: 1. Задача нагромадження інформації, 2. Задача прогнозування рекомендацій. Задача нагромадження інформації полягає в нагромадженні інформації про профілі користувачів і профілі предметів. Повнота нагромадження інформації, яка міститься в профілях користувачів і профілях предметів, впливає на точність прогнозування рекомендацій. Інформація зберігається в матриці користувач-предмет. Задача прогнозування рекомендацій полягає в прогнозуванні невідомих рейтингів предметів, які мінімізують значення у виразі (2.9). Для цього використовується інформація із профілів користувачів,які подібні до профілю активного користувача, або профілі предметів, які подібні до профілю активного предмета. Суть прогнозування рекомендацій полягає в тому, що користувачі, які мають подібні профілі, мають подібні інтереси до інших, ще невибраних предметів. Для підвищення точності прогнозування рекомендацій використовується зворотній зв’язок. Зворотній зв’язок може бути явний і неявний. Явний зворотній зв’язок передбачає корекцію прогнозованих рейтингових оцінок за

рахунок дійсних оцінок, які виставить користувач після користування предметом. При неявному зворотньому зв’язку рекомендаційна система в автоматичному режимі збирає інформацію про дії користувача. Наприклад, кількість кліків користувача по зображенню предмета, звертання користувача до повного або неповного опису характеристик предмета, історія придбання предметів користувачем. У випадку представлення профілів користувачів і предметів за допомогою атрибутів необхідний блок навчання, який формує найбільш значущі атрибути користувачів і предметів. В подальшому користувачу пропонують N предметів, які мають найбільші прогнозовані рейтинги. Переважно це десять предметів з найвищими рейтингами.

1.6.1. Рекомендаційні системи, які використовують контентну фільтрацію

Рекомендаційні системи, які використовують контентну фільтрацію, аналізують множину описів властивостей предметів, які раніше були вибрані користувачами [28–48]. На основі цього аналізу рекомендаційна система будує профілі інтересів користувачів. Профіль представляє собою структуроване надання інтересу користувача і призначений для прогнозування рекомендацій нових предметів. Процес рекомендації полягає у співставленні профілів користувачів і атрибутів профілів предметів. Результатом є оцінка рівня релевантності, яка визначає рівень інтересу користувача до предмета. Рекомендаційні системи, які використовують контентну фільтрацію, не вимагають даних про інших користувачів і можуть рекомендувати нові предмети і предмети, які не користуються великою популярністю. До недоліків контентно-орієнтованих систем можна віднести те, що вони використовують явний опис характеристик предметів і не змінюють його в процесі роботи.

1.6.2. Колаборативні рекомендаційні системи

Ідея колаборативної (спільної) фільтрації полягає в тому, що користувачі з подібними профілями (інтересами) виявляють інтерес до подібних предметів, а подібні предмети вибирають подібні користувачі. В методі колаборативної фільтрації використовується множина профілів користувачів і множина профілів предметів [49,50]. Профіль користувача і профіль предмета містять числові рейтингові оцінки, які користувачі виставили вже вибраним предметам. Числова рейтингова оцінка – це ціле додатнє число, яке може приймати значення із певної множини $\mathbf{R} = \{r_{\min}, r_{\max}\}$. При цьому $r_{\min} = 0, r_{\max} = 5$ або 10. Рейтинговою оцінкою може бути також двійкове значення $r_{like} = 1, r_{dislike} = 0$. Прогноз рекомендацій для активного предмета або активного користувача здійснюється за допомогою векторів профілів подібних користувачів із околу активного користувача або векторів профілів подібних предметів із околу активного предмета. При цьому активний користувач – це користувач, який звернувся до рекомендаційної системи для отримання рекомендацій, активний предмет – це предмет, для якого рекомендаційна система повинна здійснити прогноз оцінок користувачів. На даний час найбільш широко використовують колаборативні рекомендаційні системи. До переваг рекомендаційних систем колаборативної фільтрації можна віднести те, що вони не вимагають явного опису характеристик предметів. Такі системи використовують рейтингові числові оцінки предметів, які відповідають дійсним уподобанням користувачів. До недоліків можна віднести велику розмірність матриці користувач-предмет, неможливість прогнозувати рекомендації для нових предметів і нових користувачів.

1.6.3. Демографічні рекомендаційні системи

Демографічні рекомендаційні системи класифікують користувачів на основі їх демографічних характеристик (вік, стать, освіта, рід занять і т.д.).

В подальшому демографічні рекомендаційні системи використовують числові рейтингові оцінки користувачів, які вони виставили предметам і які зберігаються в профілях користувачів [52–56]. Демографічні рекомендаційні системи використовують для прогнозування обчислювальні вирази, які подібні до методу предмет-предмет в колаборативних рекомендаційних системах. Однак, на відміну від методу предмет-предмет в колаборативних рекомендаційних системах, в демографічних рекомендаційних системах використовують коефіцієнти подібності між векторами демографічних профілів користувачів. До переваг демографічних рекомендаційних систем можна віднести відсутність потреби в історії оцінок предметів користувачами. Демографічні рекомендаційні системи рекомендують однакові предмети для користувачів з однаковими демографічними характеристиками і не враховують рівні інтереси користувачів в однакових демографічних групах, що значно обмежує вибір користувачів.

1.6.4. Системи рекомендацій, які використовують знання

Традиційні рекомендаційні системи (контентно-орієнтовані рекомендаційні системи, колаборативні рекомендаційні системи) добре підходять для рекомендацій предметів, які володіють певними властивостями і відповідають певним уподобанням користувачів. Це книги, музика, фільми, новини. Однак, для таких предметів, як автомобілі, комп’ютери, квартири, такий підхід не є вдалим. Наприклад, певну квартиру купують не часто. Це не дозволяє нагромадити достатню кількість оцінок, які необхідні для методу колаборативної фільтрації. Окрім того, рекомендації не є вдалими, якщо вони використовують оцінки річної давнини чи старіші.

До систем рекомендацій, які використовують знання, належать системи рекомендацій на основі прецендентів CBR (case-based recommender systems) і рекомендаційні системи, які побудовані на основі вмісту факторів, які характеризують предметну область (constraint-based recommender systems) [41,57–65]. Системи рекомендацій на основі прецендентів прогнозують

рекомендації на основі метрик подібності. Рекомендаційні системи, які побудовані на основі вмісту факторів, які характеризують предметну область, прогнозують рекомендації на основі бази знань, яка містить явні правила про те, як пов’язані між собою вимоги користувачів і властивості предметів.

Рекомендації на основі прецендентів використовують методології, які застосовуються при побудові експертних систем і які базуються на накопиченому досвіді. На відміну від експертних систем, що діють на основі логічних правил, CBR-системи зберігають успішні рішення ряду реальних проблем, так звані case (приклади, або прецеденти), і при появі нової проблеми знаходять (за певним алгоритмом, найчастіше за допомогою машини логічного висновку, з кількісною оцінкою) найбільш підходящі (схожі) прецеденти, після чого пропонують відповідно модифіковану комбінацію їх рішень. Якщо нова проблема виявляється таким чином успішно вирішеною, це рішення заноситься (вводиться) в базу прецедентів для підвищення ефективності системи в майбутньому. Недолік CBR-систем в тому, що вони не створюють моделей або правил, які узагальнюють накопичений досвід.

1.6.5. Системи рекомендацій для груп користувачів

В багатьох прикладних задачах виникає потреба в прогнозуванні рекомендацій для груп користувачів [66–74]. До таких задач відносяться задачі прогнозування літературних творів певної тематики для груп користувачів бібліотек, інтернет ресурсів; рекомендації музичних творів для груп слухачів; рекомендації новин для інтернет-користувачів; рекомендації туристичних маршрутів; рекомендації товарів для груп покупців в задачах електронної комерції. Системи прогнозування рекомендацій для груп користувачів на відміну від попередніх, розглянутих вище підходів до побудови рекомендаційних систем, вимагають розв’язання певних специфічних задач, які характерні тільки для цієї предметної області. До таких задач відносяться: виділення груп користувачів з подібними уподобаннями;

агрегація уподобань групи користувачів в єдине інтегроване уподобання всієї групи користувачів; прогнозування таких рекомендацій, котрі би задовольняли всіх членів групи, тобто в групі не повинно бути користувачів, ступінь задоволення потреб яких значно відрізнявся би від ступеня задоволення потреб інших членів групи. У зв'язку із зростанням кількості користувачів Інтернету в розробленні рекомендаційних систем все більше уваги приділяється розробленню рекомендаційних систем для груп користувачів.

1.6.6. Основні проблеми в прогнозуванні рекомендацій

До основних проблем при прогнозуванні рекомендацій в рекомендаційних системах належать: проблема «холодного старту», проблема розрідженості матриці користувач-предмет, проблема масштабованості, проблема «довгого хвоста».

Проблема «холодного старту»

Проблема «холодного старту» виникає при введенні в систему нового предмета, або при зверненні до системи нового користувача [41,75–81]. При цьому профіль нового предмета і нового користувача порожні і апріорі не містять жодної інформації. Прогнозування рекомендацій для нового користувача або нового предмета при порожніх профілях неможливе, особливо в системах колаборативної фільтрації. Проблема « холодного старту » може бути вирішена різними шляхами: 1) опитуванням нових користувачів з метою отримання попередніх рейтингів деяких предметів; 2) опитування нових користувачів з метою виявлення уподобань для певних груп предметів; 3) опитування нових користувачів для отримання демографічних характеристик; 4) аналіз інформації в соціальних мережах для отримання уподобань і демографічних характеристик; 5) отримання контентної інформації про предмети із інших веб-сайтів.

Проблема розрідженості матриці користувач-предмет

Ця проблема викликана тим, що кожен конкретний користувач не може вибрати і оцінити всі предмети інтернет-магазину. У зв'язку з цим матриця користувач-предмет в методі колаборативної фільтрації містить порядка 6-7% ненульових елементів [41,57]. Це вимагає розроблення спеціальних методів зберігання інформації і алгоритмів роботи з такими структурами даних. В методі колаборативної фільтрації при прогнозуванні рекомендацій використовуються коефіцієнти подібності між векторами профілів користувачів і векторами профілів користувачів. При розрахунку коефіцієнтів подібності враховуються лише ті елементи векторів профілів, які відмінні від нульових значень в однакових елементах векторів профілів. Велика розрідженість векторів профілів приводить до того, що кількість елементів профілів, які використовуються при розрахунку коефіцієнтів подібності, менша за розрідженість самої матриці користувач-предмет. Це суттєво впливає на достовірність прогнозування рекомендацій [82]. Одним із шляхів вирішення проблеми розрідженості є використання демографічних характеристик користувачів і характеристик властивостей предметів [83], методи розкладу матриці користувач-предмет на сингулярні значення [55], використання контентної інформації в колаборативній фільтрації [84,85].

Проблема масштабованості

Проблема масштабованості полягає в здатності рекомендаційних систем обробляти великі об'єми інформації, які постійно зростають [41,57]. При цьому зростання інформації не повинно впливати на ефективність роботи рекомендаційних систем. Наприклад, американська транснаціональна компанія Amazon, яка спеціалізується на електронній комерції, здатна рекомендувати більш ніж 18 мільйонів предметів для понад 20 мільйонів

користувачів [22]. Для вирішення проблеми масштабованості використовують такі техніки: кластеризація, зменшення розмірності, мережі Байєса.

Проблема довгого хвоста»

Традиційно рекомендаційні системи пропонують користувачам предмети, які найбільш точно відповідають уподобанням активного користувача [41,57,84]. При цьому під точністю розуміється максимальна подібність вектора профілю активного користувача і векторів профілів користувачів в околі прогнозування. Такий підхід до прогнозування рекомендацій пропонує користувачу дуже подібні предмети. Рекомендаційні системи, які прогнозують користувачам предмети із множини максимально подібних профілів предметів, як правило прогнозують найбільш популярні предмети. Такі предмети в переважній більшості можуть бути вибрані користувачами самостійно. Множина таких предметів не перевищує 20% від усіх предметів, які може запропонувати інтернет-магазин. 80% предметів, це предмети, які не користуються великою популярністю, але можуть представляти інтерес для користувачів. Такі предмети належать до множини предметів «довгого хвоста». Задача сучасних рекомендаційних систем полягає в тому, щоб пропонувати користувачам не тільки популярні предмети, але і предмети із множини «довгого хвоста». Такий підхід збільшує асортимент рекомендованих товарів і тим самим впливає на дохід інтернет-магазину.

1.7. Висновки до Розділу 1

У **першому розділі** дисертаційної роботи проведено дослідження сучасного стану інформаційно-комунікаційних технологій. Виконано аналіз понять: електронний бізнес, електронна комерція, наведені характеристики основних класів електронної комерції. Показано, що інтернет-магазин є однією з найпоширеніших моделей електронної комерції в споживчому секторі (B2C). Наведені переваги і недоліки інтернет-магазинів. Розроблена структурна схема функціонування інтернет-магазину без рекомендаційних систем і з рекомендаційними системами. Виділені наступні типи рекомендаційних систем інтернет-магазину: основна рекомендаційна система, рекомендаційна система супутніх предметів, рекомендаційна система додаткових предметів. Проведено аналіз структур веб-сайтів для інтернет-магазину і показано зв'язок структур веб-сайту з типом рекомендаційної системи. Розроблені основні метрики ефективності інтернет-магаинів, які залежать від функціонування рекомендаційної системи: коефіцієнт конверсії, коефіцієнт лояльності, коефіцієнт супутніх продаж, коефіцієнт додаткових продаж. Проведено аналіз сучасних методів побудови рекомендаційних систем. Виділені наступні види рекомендаційних систем: колаборативні рекомендаційні системи; рекомендаційні системи, які використовують контентну фільтрацію; демографічні рекомендаційні системи; системи рекомендацій, які використовують знання; системи рекомендацій для груп користувачів. Показано, що на даний час найбільш широко використовуються колаборативні рекомендаційні системи. Наведена характеристика основних проблем, які виникають при розробленні рекомендаційних систем: проблема «холодного старту», проблема розрідженості матриці користувач-предмет, проблема масштабованості, проблема «довгого хвоста».

Основні положення цього розділу викладені у публікаціях автора [5, 7, 8, 9].

РОЗДІЛ 2. РОЗРОБЛЕННЯ І ДОСЛІДЖЕННЯ МОДЕЛЕЙ І МЕТОДІВ ПРОГНОЗУВАННЯ РЕКОМЕНДАЦІЙ ДЛЯ ІНТЕРНЕТ-МАГАЗИНУ

Функціонування рекомендаційної системи складається з наступних етапів (рис. 2.1) [41,57,76]:

1. нагромадження інформації про користувачів і предмети;
2. навчання;
3. прогнозування;
4. зворотний зв’язок.

Рис. 2.1 – Етапи функціонування рекомендаційної системи

На першому етапі відбувається збір і нагромадження інформації про користувачів і предмети для формування профілів користувачів і предметів. Ця інформація включає характеристики поведінки користувача, атрибути користувача, інформацію про ресурси, до яких звертається користувач,

інформацію про атрибути предметів. Для коректної роботи рекомендаційної системи потрібна найповніша інформація про користувачів і предмети. Для нагромадження достовірної інформації про уподобання і інтереси користувачів система використовує явний і неявний зворотний зв’язок.

На другому етапі використовуються методи і алгоритми навчання для обробки інформації про користувачів, яка отримується за допомогою зворотного зв’язку (рис. 2.2).

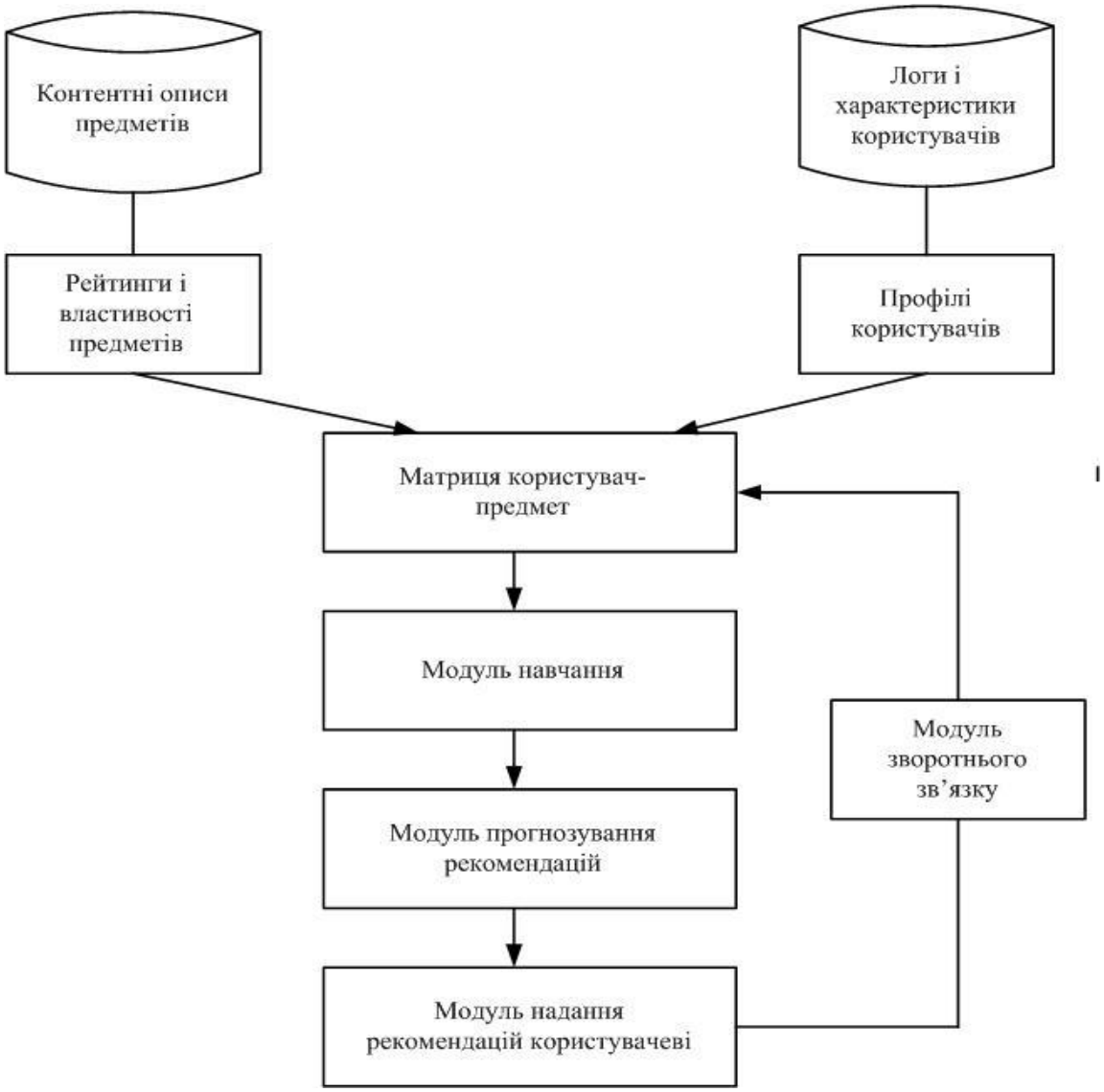


Рис. 2.2 – Узагальнена архітектура рекомендаційної системи

В рекомендаційних системах використовується зворотний зв’язок двох видів: явний і неявний. Явний зворотний зв’язок полягає в тому, що користувач коректує прогнозовані рейтинги предметів після їх використання.

При неявному зворотньому зв'язку система автоматично визначає уподобання користувача шляхом відслідковування різних дій користувача, таких, як історія покупок, історія навігації і час, який користувач проводить на веб-сторінках, посилання за якими здійснює навігацію користувач, послідовність натискання кнопок на веб-сторінках.

Узагальнена архітектура рекомендаційної системи наведена на рис. 2.2.

Нехай $\mathbf{U}$ − множина векторів профілів користувачів, $\mathbf{I}$ – множина профілів предметів, тоді множина $\mathbf{R} = \mathbf{U} \times \mathbf{I}$ − множина можливих рекомендацій рекомендаційної системи.

2.1. Модель прогнозування рекомендацій предметів інтернет-магазину методом колаборативної фільтрації

Модель прогнозування рекомендацій для користувачів і предметів інтернет-магазину полягає в наступному [41,57].

Вхідними даними для прогнозування рекомендацій предметів в інтернет-магазині є множина користувачів $\mathbf{U}$ і множина предметів $\mathbf{I}$.

Нехай $|\mathbf{U}| = m$ і $|\mathbf{I}| = n$ − кардинальні числа цих множин. Множину користувачів $\mathbf{U}$ можна подати у вигляді об'єднання двох підмножин $\mathbf{U_1}$ і $\mathbf{U_2}$:

$$\mathbf{U} = \mathbf{U_1} \cup \mathbf{U_2}, \tag{2.1}$$

$$\mathbf{U_1} \cap \mathbf{U_2} = \varnothing, \tag{2.2}$$

де $\mathbf{U_1}$ − множина користувачів, які вже здійснили придбання певних предметів через сайт інтернет-магазину;

$\mathbf{U_2}$ − множина користувачів, які заходили на сайт інтернет магазину, але ще не придбали жодного предмету і не здійснили жодних дій на сайті інтернет-магазину.

Множину предметів $\mathbf{I}$ можна подати у вигляді об'єднання двох підмножин $\mathbf{I_1}$ і $\mathbf{I_2}$:

$$\mathbf{I} = \mathbf{I_1} \cup \mathbf{I_2}, \tag{2.3}$$

$$\mathbf{I_1} \cap \mathbf{I_2} = \varnothing, \tag{2.4}$$

де $\mathbf{I_1}$ − множина предметів, які вже вибрав хоча б один користувач;

$\mathbf{I_2}$ − множина предметів, які ще не вибрав жоден користувач.

Елементами множина $\mathbf{U}$ є вектори профілів користувачів:

$$\mathbf{U} = \{\mathbf{U_{pr1}}, \mathbf{U_{pr2}}, \ldots, \mathbf{U_{prm}}\}, \tag{2.5}$$

$$\mathbf{U_{pri}} = \{r_{1i}, r_{2i}, \ldots, r_{ni}\}, \tag{2.6}$$

де $\mathbf{U_{pri}}$ − вектор профілю і-того користувача,

r_{ji} − рейтинг (оцінка) j-того предмета і-тим користувачем.

Елементами множини $\mathbf{I}$ є вектори профілів предметів:

$$\mathbf{I} = \{\mathbf{I_{pr1}}, \mathbf{I_{pr2,\ldots,}}\mathbf{I_{prm}}\}, \tag{2.7}$$

$$\mathbf{I_{prj}} = \{r_{1j}, r_{2j}, \ldots, r_{mj}\}, \tag{2.8}$$

де $\mathbf{I_{prj}}$− вектор профілю *j*-того предмета;

r_{ij} − рейтинг (оцінка) *i*-того предмета *j*-тим користувачем.

Завданням рекомендаційної системи є прогноз персоналізованих рекомендацій $\hat{r}_{ij} = Forecast(u, i, \ldots) \approx r_{ij}$, де $\hat{r}_{ij}$− прогнозоване значення оцінки, r_{ij}− дійсна оцінка, яку виставив користувач.

В результаті функціонування рекомендаційної системи потрібно досягнути мінімальної різниці між $\hat{r}_{ij}$ і r_{ij}

$$\left|\hat{r}_{ij} - r_{ij}\right| \Rightarrow \min. \tag{2.9}$$

2.2. Аналіз методів обчислення коефіцієнтів подібності векторів профілів користувачів і предметів

Традиційними методами розрахунку коефіцієнтів подібності векторів профілів користувачів і векторів профілів предметів є коефіцієнт кореляції Пірсона і косинус кута між векторами профілів. Косинусна міра подібності наведна в формулі [87]:

$$sim(\mathbf{U}, \mathbf{V}) = \frac{\sum_{i \in \mathbf{I}'} (r_{u_i} * r_{v_i})}{\sqrt{\sum_{i \in \mathbf{I}'} r_{u_i}^{\ 2} \quad * \sum_{i \in \mathbf{I}'} r_{v_i}^{\ 2}}}, \tag{2.10}$$

де $\mathbf{U}, \mathbf{V}$ − вектори профілів між якими обчислюється міра подібності;

r_{u_i} − i-та компонента вектора профілю **U**;

r_{v_i} − i-та компонента вектора профілю **V**;

$\mathbf{I}'$ − множина компонент векторів **U** і **V**, які відмінні від нуля.

Косинусна міра подібності популярна в багатьох колаборативних рекомендаційних системах для обчислення коефіцієнтів подібності між векторами профілів предметів. Вона дає незадовільний результат при наявності декількох предметів з однаковими рейтинговими оцінками. також погано враховує відмінності у величинах компонент рейтингових векторів.

Коректована косинусна міра подібності обчислюється за наступною формулою [88]:

$$sim(\mathbf{I}, \mathbf{J}) = \frac{\sum_{U \in U'} (r_{u_i} - \bar{r}_{\mathbf{I}}) * (r_{v_i} - \overline{r_{\mathbf{J}}})}{\sqrt{\sum_{U \in U'} (r_{u_i} - \bar{r}_{\mathbf{I}})^2} \sqrt{\sum_{U \in U'} (r_{v_i} - \overline{r_{\mathbf{J}}})^2}}, \tag{2.11}$$

де $\bar{r}_{\mathbf{I}}$ − середнє значення рейтингів для вектора профілю предмета **I**;

$\overline{r_{\mathbf{j}}}$ − середнє значення рейтингів для вектора профілю предмета **J**;

U' − множина користувачів, які мають спільно оцінені предмети.

Ця міра подібності дає велику похибку при малому значенні величини $|\mathbf{U}'|$. При цьому коректована косинусна міра подібності може давати велике

або мале значення при значній відмінності рейтингових величин у векторах профілів.

Коефіцієнт кореляції Пірсона обчислюється за наступною формулою [89]:

$$sim(\mathbf{U}, \mathbf{V}) = \frac{\sum_{i \in \mathbf{I}'} (r_{u_i} - \overline{r_{\mathbf{U}}})(r_{v_i} - \overline{r_{\mathbf{V}}})}{\sqrt{\sum_{i \in \mathbf{I}'} (r_{u_i} - \overline{r_{\mathbf{U}}})^2} \sqrt{\sum_{i \in \mathbf{I}'} (r_{v_i} - \overline{r_{\mathbf{V}}})^2}} \ . \quad (2.12)$$

Коефіцієнт кореляції Пірсона лежить в межах [-1, 1]. Значення +1 вказує на велику позитивну кореляційну залежність. Значення -1 вказує на велику негативну кореляційну залежність. Коефіцієнт кореляції Пірсона дає неправильні значення, якщо в обидвох векторах профілів є декілька однакових значень рейтингових оцінок. Він також обчислює неправильні значення при малому значенні величини $|\mathbf{I}'|$. Цей коефіцієнт може мати велику або низьку величину подібності незважаючи на подібність або відмінність рейтингових оцінок.

Проста рекомендація є найбільш простим видом рекомендації. Вона використовується у тому випадку, коли необхідно отримати мінімальний час розрахунку. Отриманий рейтинг є середнім арифметичним значенням рейтингів у векторі - рядку або векторі - стовпці матриці **A**. Проста рекомендація має таку формулу розрахунку [22]:

$$r'_{ij} = \frac{\sum_{j=1}^{n} r_{ij}}{n} \ . \quad (2.13)$$

Евклідова відстань найбільш поширена функція відстані між векторами. Представляє собою геометричну відстань в багатовимірному просторі [90]:

$$d(\bar{x}, \bar{y}) = \sqrt{(x_i - y_i)^2} \ . \quad (2.14)$$

Манхеттенівська відстань (відстань міських кварталів) − ця сума абсолютних величин різниць координат векторів [90]. У більшості випадків ця міра відстані призводить до таких же результатів, що і звичайна відстань Евкліда. Однак для цього підходу вплив окремих великих різниць (викидів)

зменшується (тому що вони не зводяться в квадрат). Формула для розрахунку манхеттенівської відстані:

$$d(\bar{x}, \bar{y}) = \sum_{i=1}^{n} |x_i - y_i| \ . \tag{2.15}$$

Умовний (обмежений) коефіцієнт кореляції Пірсона обчислюється за формулою [91]:

$$sim(\mathbf{U}, \mathbf{V}) = \frac{\sum_{i \in \mathbf{I}'} (r_{u_i} - r_{med})(r_{v_i} - r_{med})}{\sqrt{\sum_{i \in \mathbf{I}'} (r_{u_i} - r_{med})^2} \sqrt{\sum_{i \in \mathbf{I}'} (r_{v_i} - r_{med})^2}} \ , \tag{2.16}$$

де r_{med} − медіана впорядкованої множини можливих значень рейтингових оцінок.

Ця оцінка подібності вносить незначне покращення в звичайний коефіцієнт кореляції Пірсона. Він також обчислює неправильні значення при наявності навіть незначної кількості однакових рейтингових оцінок в обидвох векторах профілів.

Середня квадратична різниця обчислюється за наступною формулою [91]:

$$sim(\mathbf{U}, \mathbf{V}) = 1 - \frac{\sum_{i \in \mathbf{I}'} (r_{u_i} - r_{v_i})^2}{|\mathbf{I}'|} . \tag{2.17}$$

Цей коефіцієнт подібності не враховує однакові значення рейтингових оцінок в векторах профілів користувачів.

Коефіцієнт подібності Жакарда обчислюється за наступною формулою [92]:

$$sim(\mathbf{U}, \mathbf{V}) = \frac{|\mathbf{I_u} \cap \mathbf{I_v}|}{|\mathbf{I_u} \cup \mathbf{I_v}|} \ , \tag{2.18}$$

де $\mathbf{I_u}$ − множина предметів, які вибрав користувач **U**;

$\mathbf{I_v}$ − множина предметів, які вибрав користувач **V**.

Ця міра подібності не враховує абсолютні значення величин рейтингових оцінок.

Жакард+середня квадратична різниця обчислюється за наступною формулою [93]:

$$sim(\mathbf{U}, \mathbf{V}) = \frac{|\mathbf{I_u} \cap \mathbf{I_v}|}{|\mathbf{I_u} \cup \mathbf{I_v}|} (1 - \frac{\sum_{i \in \mathbf{I}'} (r_{u_i} - r_{v_i})^2}{|\mathbf{I}'|}). \quad (2.19)$$

В цій мірі подібності присутні ті ж недоліки, що і у коефіцієнті подібності Жакарда і середній квадратичній різниці.

В загальному для розглянутих вище мір подібності можна виділити наступні недоліки:

1. Ці міри подібності погано працюють при невеликій кількості спільно оцінених предметів, Подібність предметів може бути виявлена високою, якщо немає спільних користувачів, які б оцінили ці предмети. Наприклад, нехай задані профілі предметів $\mathbf{I} = (1,0,2,0,1,0,2,0,3,0)$ і $\mathbf{J} = (0,1,0,2,0,1,0,3)$. Ці вектори профілів містять однакові рейтингові оцінки, однак розглянуті вище міри подібності не зможуть обчислити їхню подібність;
2. Ці міри подібності дають невірний результат у випадку лише одного спільно оціненого предмета;
3. Ці міри подібності використовуються для обчислення лише локальної подібності в околі активного предмета або користувача.

2.3. Метод користувач-користувач прогнозування рейтингів

Отримання коефіцієнтів близкості векторів для розрахунку рекомендації є важливим кроком в системі колаборативного фільтрування (КФ). У КФ алгоритмі на основі найближчих сусідів підмножина найближчих сусідів для активного користувача вибирається з урахуванням їх схожості з ним і зважену сукупність їх оцінок використовують для генерування прогнозу для активного користувача [41,57,93].

Обчислення прогнозу і-того предмету для активного користувача здійснюється за наступною формулою:

$$P_{a,i} = \bar{r}_a + \frac{\sum_{u \in \mathbf{U}} (r_{ui} - \bar{r}_u) w_{au}}{\sum_{u \in \mathbf{U}} |w_{au}|} \ , \qquad (2.20)$$

де $\bar{r}_a$ і $\bar{r}_u$ − середні оцінки для вектора рейтингів активного користувача *a* і користувача *u*;

w_{au} − ваговий коефіцієнт близькості вектора рейтингів активного користувача *a* і користувача *u*;

U − множина векторів-рейтингів користувачів, які мають спільно оцінені предмети.

2.4. Метод предмет-предмет прогнозування рейтингів

Для прогнозу на основі предметів використовується проста зважена середня величина [88]:

$$P_{ui} = \frac{\sum_{n \in N} r_{un} w_{in}}{\sum_{n \in N} |w_{in}|} \ . \qquad (2.21)$$

2.5. Метод розрахунку коефіцієнту подібності з урахуванням розрідженості і довжини векторів профілів

Даний метод розроблений здобувачем в роботі [6]. В [6] показано, що міри подібності між векторами, які ґрунтуються на косинусі кута між векторами і коефіцієнті кореляції Пірсона можуть давати значну похибку. Дослідимо це на реальних прикладах. Нехай задано два вектори, вектор **A**(2; 3,464) і вектор **B**(3,464; 2). Довжини обидвох векторів однакові $\|\mathbf{A}\| = \|\mathbf{B}\| = 4$ кут між векторами $\alpha = 30^0$ (рис.2.3).

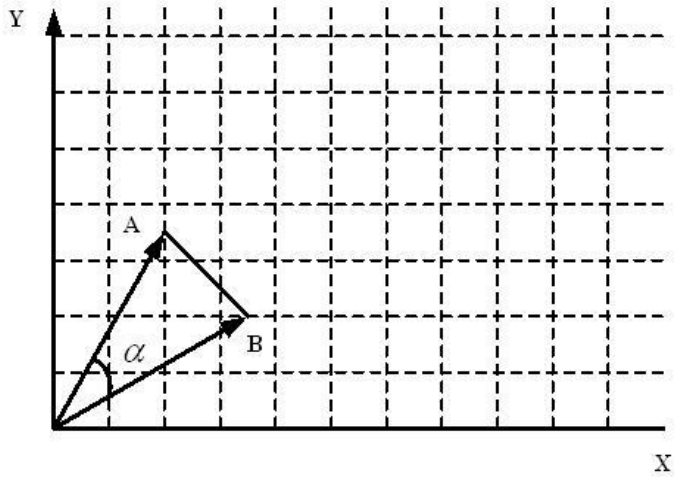


Рис. 2.3 – Тестовий приклад $\|\mathbf{A}\|$=4, $\|\mathbf{B}\|$=4, $\|\mathbf{A}-\mathbf{B}\|$=2,707, α=$30^0$

Тоді косинус кута α між векторами обчислюється за допомогою наступного виразу:

$$\cos(\alpha)=\frac{\mathbf{A_x}*\mathbf{B_x}+\mathbf{A_y}*\mathbf{B_y}}{\|\mathbf{A}\| \ * \ \|\mathbf{B}\|}=\frac{2*3{,}464+3{,}464*2}{4*4}=\frac{\sqrt{3}}{2}=0{,}866\ . \qquad (2.22)$$

Евклідова відстань між векторами обчислюється за допомогою наступного виразу:

$$D_1=\|\mathbf{A}-\mathbf{B}\|=\sqrt{(\mathbf{A_x}-\mathbf{B_x})^2+(\mathbf{A_y}-\mathbf{B_y})^2}= \sqrt{(2-3{,}464)^2+(3{,}464-2)^2}=2{,}707\ . \qquad (2.23)$$

Нехай задано два вектори, вектор **A**(4; 6,928) і вектор **B**(3,464; 2), $\|\mathbf{A}\|$=8, $\|\mathbf{B}\|$=4, кут між векторами α=30^0 (рис.2.4). Тоді косинус кута α між векторами обчислюється за допомогою наступного виразу:

$$\cos(\alpha)=\frac{\mathbf{A_x}*\mathbf{B_x}+\mathbf{A_y}*\mathbf{B_y}}{\|\mathbf{A}\| \ * \ \|\mathbf{B}\|}=\frac{4*3{,}464+6{,}928*2}{8*4}=\frac{\sqrt{3}}{2}=0{,}866\ , \qquad (2.24)$$

Евклідова відстань між векторами обчислюється за допомогою наступного виразу:

$$D_2 = \|\mathbf{A}-\mathbf{B}\| = \sqrt{(\mathbf{A_x}-\mathbf{B_x})^2+(\mathbf{A_y}-\mathbf{B_y})^2} = \sqrt{(4-3{,}464)^2+(3{,}464-2)^2} = 4{,}957 . \quad (2.25)$$

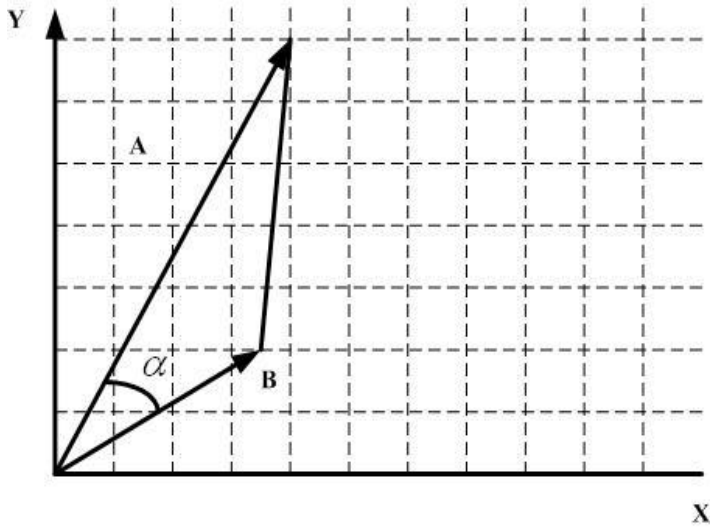


Рис. 2.4 – Тестовий приклад $\|\mathbf{A}\|$=8, $\|\mathbf{B}\|$=4, α=30^0

Нехай задано два вектори, вектор **A**(4; 6,928) і вектор **B**(6,928; 4), $\|\mathbf{A}\|$=8, $\|\mathbf{B}\|$=8, кут між векторами α=30^0 (рис.2.5). Тоді косинус кута α між векторами обчислюється за допомогою наступного виразу:

$$\cos(\alpha) = \frac{\mathbf{A_x}*\mathbf{B_x}+\mathbf{A_y}*\mathbf{B_y}}{\|\mathbf{A}\| * \|\mathbf{B}\|} = \frac{4*6{,}928+6{,}928*4}{8*8} = \frac{\sqrt{3}}{2} = 0{,}866 . \quad (2.26)$$

Евклідова відстань між векторами обчислюється за допомогою наступного виразу:

$$D_3 = \|\mathbf{A}-\mathbf{B}\| = \sqrt{(\mathbf{A_x}-\mathbf{B_x})^2+(\mathbf{A_y}-\mathbf{B_y})^2} = \sqrt{(4-6{,}928)^2+(6{,}928-4)^2} = 4{,}141 . \quad (2.27)$$

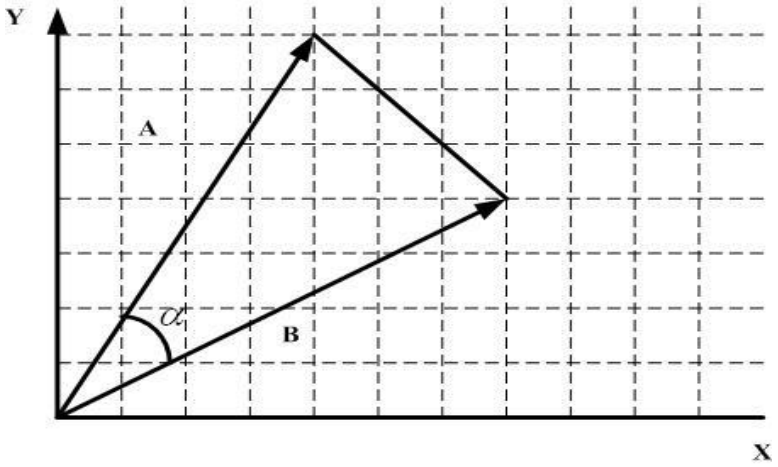


Рис. 2.5 – Тестовий приклад $\|\mathbf{A}\|$=8, $\|\mathbf{B}\|$=8, α=30^0

Нехай задано два вектори $\|\mathbf{A}\|=\|\mathbf{B}\|=8$, кут між векторами α=$0^0$, кут між векторами і віссю абсцих β=60^0 (рис. 2.6). Тоді косинус кута α між векторами обчислюється становить $\cos(\alpha)=1$. Евклідова відстань між векторами обчислюється за допомогою наступного виразу:

$$D_4=\|\mathbf{A}-\mathbf{B}\|=\|\mathbf{A}\|-\|\mathbf{B}\|=8-8=0\ . \qquad (2.28)$$

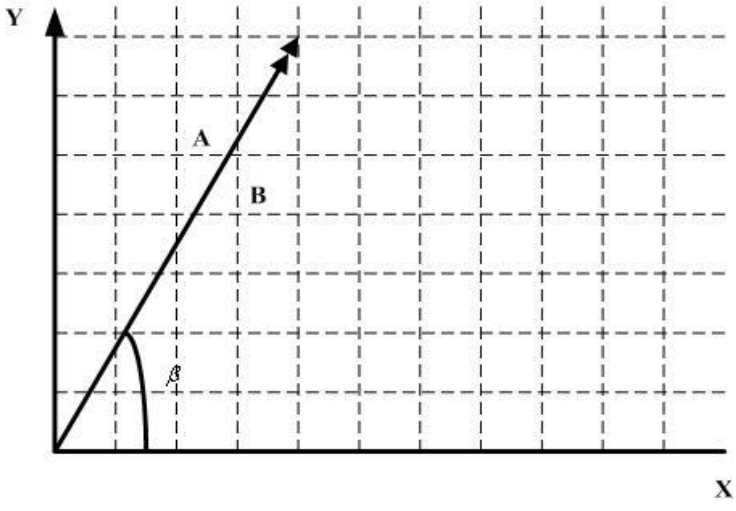


Рис. 2.6 – Тестовий приклад $\|\mathbf{A}\|=\|\mathbf{B}\|$=8, α=$0^0$, β=60^0

Нехай задано два вектори $\|\mathbf{A}\|$=8, $\|\mathbf{B}\|$=4, кут між векторами α=0^0 , кут між векторами і віссю абсцих β=60^0 (рис. 2.7). Тоді косинус кута α між векторами

обчислюється становить $\cos(\alpha)=1$. Евклідова відстань між векторами обчислюється за допомогою наступного виразу:

$$D_5 = \|\mathbf{A}-\mathbf{B}\| = \|\mathbf{A}\| - \|\mathbf{B}\| = 8-4=4\ . \qquad (2.29)$$

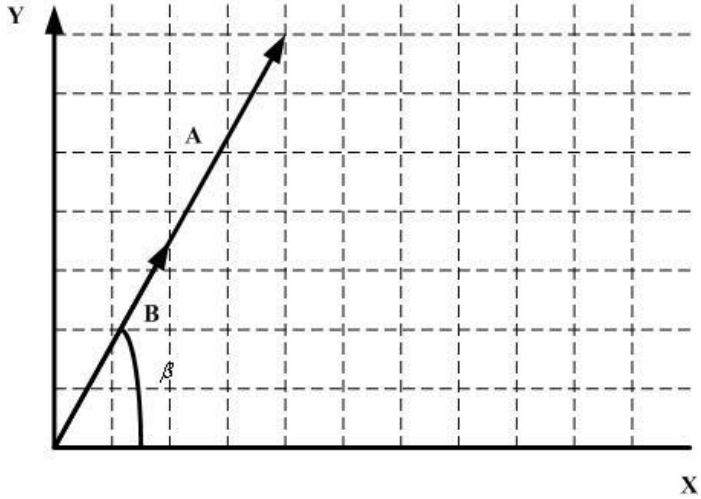


Рис. 2.7 – Тестовий приклад $\|\mathbf{A}\|$=8, $\|\mathbf{B}\|$=4, α=0^0, β=60^0

Аналіз розглянутих прикладів показує, що у першому тестовому прикладі і у другому тестовому прикладі кут між векторами α=30^0, однак D_1 < D_2. У третьому тестовому прикладі α=30^0, однак $D_2 > D_1$ і $D_2 > D_3$ У четвертому тестовому прикладі D_4 =0, $\cos(\alpha)=1, \alpha=0$. У п'ятому тестовому прикладі D_5=4, $\cos(\alpha)=1, \alpha=0.$ Більш адекватною оцінкою міри подібності між векторами буде обернена евклідова відстань:

$$sim(\mathbf{A},\mathbf{B}) \propto \frac{1}{D}\ . \qquad (2.30)$$

Для врахування випадку коли D=0 використовується наступний вираз:

$$sim(\mathbf{A},\mathbf{B}) = \frac{1}{1+D}\ . \qquad (2.31)$$

Розглянемо застосування цього виразу для наведених вище прикладів. Для прикладу наведеному на рис.2.3 розрахунок міри подібності здійснюється за допомогою наступного виразу:

$$sim_1(\mathbf{A}, \mathbf{B}) = \frac{1}{1+D_1} = \frac{1}{1+2{,}707} = 0{,}27 \,. \qquad (2.32)$$

Для прикладу наведеному на рис.2.4 розрахунок міри подібності здійснюється за допомогою наступного виразу:

$$sim_2(\mathbf{A}, \mathbf{B}) = \frac{1}{1+D_2} = \frac{1}{1+4{,}957} = 0{,}167 \,. \qquad (2.33)$$

Для прикладу наведеному на рис.2.5 розрахунок міри подібності здійснюється за допомогою наступного виразу:

$$sim_3(\mathbf{A}, \mathbf{B}) = \frac{1}{1+D_3} = \frac{1}{1+4{,}141} = 0{,}194 \,. \qquad (2.34)$$

Для прикладу наведеному на рис.2.6 розрахунок міри подібності здійснюється за допомогою наступного виразу:

$$sim_4(\mathbf{A}, \mathbf{B}) = \frac{1}{1+D_4} = \frac{1}{1+0{,}0} = 1 \,. \qquad (2.35)$$

Для прикладу наведеному на рис.2.7 розрахунок міри подібності здійснюється за допомогою наступного виразу:

$$sim_5(\mathbf{A}, \mathbf{B}) = \frac{1}{1+D_5} = \frac{1}{1+4{,}0} = 0{,}2 \,. \qquad (2.36)$$

З наведених вище розрахунків вбачається, що вираз (2.31) дає дійсну оцінку міри подібності векторів.

Нехай $\mathbf{X}, \mathbf{Y} \in R^n$ – два вектора в n-вимірному просторі. Коефіцієнт кореляції Пірсона між цими двома векторами може бути записаний наступним чином [94]:

$$r_{XY} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\ \sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} \,. \qquad (2.37)$$

Нехай $\mathbf{X}_1 = k\mathbf{X}$, тоді:

$$r_{kXY} = \frac{n\sum_{i=1}^{n} kx_i y_i - \left(\sum_{i=1}^{n} kx_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} k^2 x_i^2 - \left(\sum_{i=1}^{n} kx_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} =$$

$$\frac{k\left(n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\right)}{\sqrt{k^2\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} = r_{xy}. \quad (2.38)$$

Таким чином коефіцієнт кореляції Пірсона інваріантний стосовно масштабування. Для двох колінеарних векторів, як і для випадку косинусної міри подібності, коефіцієнт кореляції Пірсона дасть однаковий результат.

Нехай задані три вектори:

$\mathbf{A} = (1,0,1,2), \mathbf{B} = (8,0,8,16), \mathbf{C} = (5,3,5,7).$

Тоді:

$$\mathbf{B} = k\mathbf{A} = 8\mathbf{A}, \quad (2.39)$$

$$\mathbf{C} = 2\mathbf{A} + 3, \quad (2.40)$$

$$r_{AB} = r_{AC} = 1. \quad (2.41)$$

Можна зробити наступний висновок − найбільш прийнятною величиною міри подібності між векторами є обернена величина евклідової відстані між векторами.

2.6. Використання демографічних характеристик користувачів при прогнозуванні рекомендацій

Матриця користувач - предмет містить 6-7% ненульових елементів. Одна із проблем сучасних рекомендаційних систем – це проблема нового користувача. Новий користувач при входженні в рекомендаційну систему має порожній вектор профілю (вектор профілю містить лише нульові елементи).

Перспективним методом вирішення цих задач є використання демографічної інформації про користувачів [2,11]. Демографічну інформацію про користувача можна отримати при реєстрації користувача в системі, із соціальних мереж, із аналізу контенту інших сайтів. На основі аналізу демографічної інформації формується демографічний вектор профілю користувача

$$\mathbf{U_{dem}} = \{вік, стать, рід занять\}. \quad (2.42)$$

Демографічний вектор профілю можна ефективно використати для покращання точності розрахунку коефіцієнтів подібності. Для цього компоненти вектора (2.42) категоризують за наступною схемою:

Таблиця 2.1 – Таблиця категоризації компонент вектора демографічного профілю користувача

Вік				Стать		Рід занять		
вік<18	18<вік<30	30<вік<50	50<вік	Чол.	жін	лікар	інженер	.....

Всього категоризований вектор демографічного профілю має 27 компонент. Компонента категоризованого вектора профілю користувача містить 1 в тій позиції, яка відповідає категорії демографічної характеристики і 0 у решті позицій.

Після категоризації компоненти вектора (2.42) мають двійкове бітове значення. Незважаючи на те, що розмірність вектора зростає до 27, двійковий бітовий вміст вектора дозволяє ефективно визначати подібність між

категоризованими демографічними векторами профілів користувачів. Для цього використовується коефіцієнт подібності Жакарда:

$$J = \frac{\left|\mathbf{A} \cap \mathbf{B}\right|}{\left|\mathbf{A} \cup \mathbf{B}\right|}\ , \tag{2.43}$$

де **A** і **B** – це вектори, які можуть містити довільні дійсні значення або алфавітно-цифрову інформацію.

Для двійкових векторів формула розрахунку коефіцієнта подібності має наступний вигляд [90]:

$$J = \frac{M_{11}}{M_{10} + M_{01} + M_{11}}\ , \tag{2.44}$$

де M_{11} - загальна кількість елементів, де компоненти векторів **A** і **B** мають значення 1;

M_{10} - загальна кількість елементів, де компоненти вектора A мають значення 1 і компоненти вектора B мають значення 0;

M_{01} - загальна кількість елементів, де компоненти вектора A мають значення 0 і компоненти вектора B мають значення 1;

Кожен компонент векторів **A** і **B** повинен знаходитись в одній із чотирьох категорій:

$$M_{11} + M_{10} + M_{01} + M_{00} = n\ , \tag{2.45}$$

де n − розмірність векторів **A** і **B**;

M_{00} - загальна кількість елементів, де компоненти векторів **A** і **B** мають значення 0.

Обчислення коефіцієнтів у виразі (2.44) потребує виконання операцій над двійковим вмістом демографічних векторів профілів, що потребує значно

менших часових затрат, ніж обчислення коефіцієнтів подібності для рейтингових векторів профілів користувачів.

Для прогнозування рекомендацій у виразах (2.20) і (2.21) використовується модифікований коефіцієнт подібності:

$$w'_{ij} = sim(i, j) + J_{ij} \times sim(i, j) \ , \tag{2.46}$$

$$w'_{ij} = J_{ij} + J_{ij} \times sim(i, j) \ , \tag{2.47}$$

де $sim(i, j)$ − коефіцієнт подібності між рейтинговими векторами профілів користувачів;

J_{ij} − коефіцієнт подібності між демографічними векторами профілів користувачів.

Вираз (2.46) враховує домінуюче значення подібності користувачів за рейтинговими векторами профілів. Вираз (2.47) враховує домінуюче значення подібності користувачів за демографічними векторами профілів.

Для обчислення коефіцієнтів подібності між рейтинговими векторами профілів користувачів використовується метод, який базується на оберненій величині евклідової відстані між векторами.

Запропонований метод дозволяє вирішити проблему нового користувача. Для нового користувача вираз (2.47) буде мати наступний вигляд:

$$w'_{ij} = J_{ij} \ . \tag{2.48}$$

В подальшому отримані коефіцієнти подібності використовуються при прогнозуванні рекомендацій у виразах (2.20) і (2.21).

2.7. Висновки до Розділу 2

У **другому розділі** дисертаційної роботи досліджується метод зваженої суми для персоналізованого прогнозування рейтингів користувачів і предметів в колаборативних рекомендаційних системах. Показано, що на точність прогнозування істотно впливають коефіцієнти подібності векторів профілів користувачів і профілів предметів. Досліджуються наступні методи обчислення коефіцієнтів подібності: косинусна міра подібності; коректована косинусна міра подібності; коефіцієнт кореляції Пірсона; умовний (обмежений) коефіцієнт кореляції Пірсона; середня квадратична різниця; коефіцієнт подібності Жакарда; Жакард+середня квадратична різниця. Аналізуються недоліки цих мір подібності. Наведено особливості використання цих мір подібності у розрахункових виразах методу користувач-користувач і методу предмет-предмет в колаборативних рекомендаційних системах. В основному для розрахунку коефіцієнтів подібності використовується косинусна міра подібності або коефіцієнт кореляції Пірсона. Показано, що косинусна міра подібності і коефіцієнт кореляції Пірсона дають значні похибки. На конкретних прикладах доведено, що більш точною мірою подібності є обернена міра Евкліда. Вказано, що значна розрідженість матриці користувач-предмет вимагає врахування демографічних характеристик користувачів при обчисленні коефіцієнтів подібності векторів профілів користувачів і якісних характеристик предметів при обчисленні коефіцієнтів подібності векторів профілів предметів. Розроблено метод розрахунку коефіцієнтів подібності векторів профілів, який використовує демографічні характеристики користувачів. Показано, що даний метод може використовуватися для розрахунку коефіцієнтів подібності в рішенні проблеми «холодного старту».

Основні положення цього розділу викладені у публікаціях автора [2, 6, 11].

РОЗДІЛ 3. РОЗРОБЛЕННЯ І ДОСЛІДЖЕННЯ ГІБРИДНИХ МЕТОДІВ І ЗАСОБІВ ДЛЯ ПРОГНОЗУВАННЯ РЕКОМЕНДАЦІЙ В РЕКОМЕНДАЦІЙНІЙ СИСТЕМІ ДЛЯ ІНТЕРНЕТ-МАГАЗИНУ

Розвиток сучасних рекомендаційних систем вимагає підвищення ефективності аналізу та пошуку інформації у великих сховищах даних, мережах, Інтернеті [95−101]. Тому все більш важливе значення в розвитку таких систем займають перспективні інформаційні технології інтелектуального аналізу даних (Data mining) [102,103], аналізу текстової інформації (Text mining) [104], виявлення і пошуку залежностей і знань у веб-даних (Web mining) [105,106]. У кожній з цих технологій використовуються методи і алгоритми виділення груп і класів об'єктів. Серед них важливе місце займають методи та алгоритми кластерного аналізу, методи і алгоритми пошуку асоціативних правил. Важливою особливістю цих методів є те, що вони не вимагають початкових знань про поділ об'єктів на класи і групи і можуть використовуватися на початкових стадіях аналізу даних. Гібридні моделі і методи прогнозування рекомендацій можуть поєднувати декілька моделей і методів, які були описані в розділі 1.6.

3.1. Прогнозування рекомендацій на основі методу пошуку асоціативних правил

Застосування асоціативних правил в рекомендаційних системах направлене на підвищення точності прогнозування рекомендацій, вирішенні проблеми нового користувача і нового предмета, підвищення рівня захисту рекомендаційної системи від зовнішніх атак [107]. В даному розділі розроблений метод прогнозування рекомендацій, який використовує асоціативні правила і враховує рівень інтересу користувачів до попередньо вибраних предметів [4].

Асоціативні правила дозволяють знаходити закономірності між пов'язаними подіями [108,109]. Прикладом такої закономірності служить

правило, яке вказує, що з подією **X** слідує подія **Y** з деякою ймовірністю. Встановлення таких залежностей дає можливість знаходити дуже прості і інтуїтивно зрозумілі правила. Приклади застосування асоціативних правил в інткернет-магазинах:

1. Аналіз ринкового кошика (market basket analysis). Алгоритм виявляє типові шаблони покупок і спільно придбаних предметів. Отримані результати дозволяють оптимізувати асортимент предметів, збільшити обсяги продажів за рахунок пропозиції клієнтам супутніх предметів.

2. Крос-продажі (cross-sell) і продажі з підвищенням ціни (up-selling). Алгоритм дозволяє на підставі шаблонів споживчої поведінки споживачів запропонувати їм подібні предмети однієї категорії або запропонувати предмети з інших категорій з підвищенням ціни.

3. Директ мейл (direct mail) - пряма адресна розсилка рекламних пропозицій потенційним і існуючим покупцям - є високоефективним, простим і дешевим маркетинговим інструментом. Для збільшення кількості відгуків на листи необхідно проводити ретельний відбір об'єктів для розсилки послань, чому сприяє розглянутий алгоритм.

Позначимо через $\hat{\mathbf{T}}$ − множину транзакцій

$$\hat{\mathbf{T}} = (\mathbf{T_1}, \mathbf{T_2}, \ldots, \mathbf{T_k}) , \tag{3.1}$$

де $\mathbf{T_i}$ − вектор i− тої транзакції.

$$\mathbf{T_i} = (i_{1i}, i_{2i}, \ldots, i_{mi}) , \tag{3.2}$$

де i_{ji} − j− тий предмет i− тої транзакції.

Позначимо через $\mathbf{I} = (i_1, i_2, \ldots i_n)$ − множину предметів, які входять у всі транзакції. Асоціативним правилом називається імплікація $\mathbf{X} \Rightarrow \mathbf{Y}$, де $\mathbf{X} \subset \mathbf{I}, \mathbf{Y} \subset \mathbf{I}, \mathbf{X} \cap \mathbf{Y} = \varnothing$. Асоціативні правила характеризуються підтримкою supp (support) і достовірністю conf (confidence). Нехай $\mathbf{F} = \mathbf{X} \cup \mathbf{Y}$, $\mathbf{D_F}$ − множина транзакцій, в які входить $\mathbf{F}$. Тоді підтримка обчислюється за допомогою наступного виразу:

$$\text{Supp}\,(\mathbf{X}\cup\mathbf{Y}) = \frac{|\mathbf{D_F}|}{|\mathbf{T}|}\ , \tag{3.3}$$

де $|\mathbf{D_F}|$ − кардинальне число множини $\mathbf{D_F}$,

$|\mathbf{T}|$ − кардинальне число множини транзакцій **T**.

Підтримка показує яку частину складає кількість транзакцій, які містять набір даних $\mathbf{X}\cup\mathbf{Y}$ від загальної кількості транзакцій.

Достовірність обчислюється за наступною формулою:

$$\text{Conf}\ (\mathbf{X}\Rightarrow\mathbf{Y}) = \frac{\sup p(\mathbf{X}\cup\mathbf{Y})}{\sup p(\mathbf{X})}\ . \tag{3.4}$$

Достовірність показує яка ймовірність того, що при наявності в транзакції набору **X** в ній також присутній набір**Y**. Чим більша достовірність, тим краще правило.

Підтримка і достовірність − це два основних параметри, які визначають корисність правила. Однак ці два параметри не дозволяють повністю оцінити корисність правила. Якщо виконується умова:

$$\text{Conf}(\mathbf{X}\Rightarrow\mathbf{Y}) = \frac{\text{Supp}(\mathbf{X}\cup\mathbf{Y})}{\text{Supp}(\mathbf{X})} < \text{Supp}(\mathbf{Y})\ , \tag{3.5}$$

це означає, що ймовірність випадково вгадати наявність в транзакції набору **Y** є більшою, ніж ймовірність появи цього набору в асоціативному правилі $\mathbf{X}\Rightarrow\mathbf{Y}$. Високі рівні підтримки і достовірності самі по собі ще не свідчать про значимість виявленої асоціації. Предмети можуть бути дуже популярними і тільки тому часто зустрічаються в одній транзакції. Якщо рішення про вибір двох предметів незалежні, тоді говорити про якесь правило, що їх пов'язує, не доводиться. З математичної статистики відомо, що якщо умова і наслідок не залежать один від одного, тоді підтримка правила в цілому буде приблизно дорівнює добутку тільки підтримки лише умови і підтримки лише наслідку $\text{Sup}(\mathbf{X}\Rightarrow\mathbf{Y})\approx\text{Sup}(\mathbf{X})\text{Sup}(\mathbf{Y})$.

Для більш точної оцінки корисності правила вводиться наступна міра – поліпшення (improvement):

$$\text{Impr}(\mathbf{X} \Rightarrow \mathbf{Y}) = \frac{\text{Supp}(\mathbf{X} \cup \mathbf{Y})}{\text{Supp}(\mathbf{X})\,\text{Supp}(\mathbf{Y})} \ . \qquad (3.6)$$

Поліпшення - це відношення частоти появи наслідку в транзакціях, які також містять і умову, до частоти появи наслідку в цілому. Тому, якщо $\text{Impr} > 1$, більш вірогідна поява наслідку у транзакціях, що містять умову, ніж у всіх інших. Можна сказати, що Impr є узагальненою мірою зв'язку двох предметних наборів: при $\text{Impr} > 1$ зв'язок позитивний, при $\text{Impr} = 1$ він відсутній, а при $\text{Impr} < 1$ - негативний.

Однакову величину поліпшення можуть мати два асоціативних правила. Наприклад, нехай $\text{Supp}(\mathbf{X} \cup \mathbf{Y}) = 0{,}5, \text{Supp}(\mathbf{X}) = 0{,}5, \text{Supp}(\mathbf{Y}) = 0{,}5$. Тоді достовірність має наступну величину:

$$\text{Impr}(\mathbf{X} \Rightarrow \mathbf{Y}) = \frac{\text{Supp}(\mathbf{X} \cup \mathbf{Y})}{\text{Supp}(\mathbf{X})\,\text{Supp}(\mathbf{Y})} = \frac{0{,}5}{0{,}5\;0{,}5} = 2 \ . \qquad (3.7)$$

В іншому прикладі $\text{Supp}(\mathbf{X} \cup \mathbf{Y}) = 0{,}8, \text{Supp}(\mathbf{X}) = 0{,}5, \text{Supp}(\mathbf{Y}) = 0{,}8$. Тоді достовірність має наступну величину:

$$\text{Impr}(\mathbf{X} \Rightarrow \mathbf{Y}) = \frac{\text{Supp}(\mathbf{X} \cup \mathbf{Y})}{\text{Supp}(\mathbf{X})\,\text{Supp}(\mathbf{Y})} = \frac{0{,}8}{0{,}5\;0{,}8} = 2 \ . \qquad (3.8)$$

Для врахування таких випадків вводиться величина підсилення:

$$\text{Leverage}(\mathbf{X} \cup \mathbf{Y}) = \left|\text{Supp}(\mathbf{X} \cup \mathbf{Y}) - \text{Supp}(\mathbf{X})\,\text{Supp}(\mathbf{Y})\right| \ . \qquad (3.9)$$

Асоціативне правило тим краще, чим більша величина $\text{Leverage}(\mathbf{X} \cup \mathbf{Y})$.

Метою аналізу асоціативних правил є встановлення наступних залежностей: якщо в транзакції зустрівся деякий набір елементів **X**, то на підставі цього можна зробити висновок про те, що інший набір елементів **Y** також повинен з'явитися в цій транзакції. Встановлення таких залежностей дає можливість знаходити дуже прості і інтуїтивно зрозумілі правила.

Алгоритми пошуку асоціативних правил призначені для знаходження всіх правил $\mathbf{X} \Rightarrow \mathbf{Y}$, причому підтримка і достовірність цих правил повинні бути вище деяких наперед визначених порогових значень, які називаються

відповідно мінімальною підтримкою (minsupport) і мінімальною достовірністю (minconfidence).

Завдання знаходження асоціативних правил розбивається на дві підзадачі:

1. Знаходження всіх наборів елементів, які задовольняють порогу minsupport. Такі набори елементів називаються популярними;

2. Генерація правил з популярних наборів елементів з достовірністю, що задовольняє порогу minconfidence.

Для цього використовується алгоритм Apriori [108,109].

Значення для параметрів мінімальна підтримка і мінімальна достовірність вибираються таким чином, щоб обмежити кількість знайдених правил. Якщо підтримка має велике значення, то алгоритм буде знаходити правила, які добре відомі аналітикам або настільки очевидні, що немає ніякого сенсу проводити такий аналіз. З іншого боку, низьке значення підтримки веде до генерації величезної кількості правил, що, звичайно, вимагає істотних обчислювальних ресурсів. Тим не менше, більшість цікавих правил знаходиться саме при низькому значенні порогу підтримки. Хоча занадто низьке значення підтримки веде до генерації статистично необґрунтованих правил.

Пошук асоціативних правил зовсім не тривіальне завдання, як може здаватися на перший погляд. Одна з проблем - алгоритмічна складність при знаходженні частих зустрічних наборів елементів, тому що з ростом числа елементів в $\mathbf{I}$ ($|\mathbf{I}|$) експоненціально зростає число потенційних наборів елементів.

Для того, щоб було можливо застосувати алгоритм Apriori, необхідно провести попередню обробку даних: по-перше, привести всі дані до бінарного значення; по-друге, змінити структуру даних.

Нехай задана база даних транзакцій у вигляді таблиці 3.1.

Таблиця 3.1 – База даних транзакцій

Номер транзакції	Найменування елемента	Кількість
1001	A	2
1001	D	3
1001	E	1
1002	A	2
1002	F	1
1003	B	2
1003	A	2
1003	C	2
…..	…..	…..

Таблиця 3.2 – Нормалізоване представлення бази даних транзакцій

TID	A	B	C	D	E	F	G	H	I	K	….
1001	1	0	0	1	1	0	0	0	0	0	….
1002	1	0	0	0	0	1	0	0	0	0	….
1003	1	1	1	0	0	0	0	0	1	0	….

Нормалізоване представлення бази даних представлене в табл. 3.2. Кількість стовпців в таблиці дорівнює кількості елементів, присутніх в множині транзакцій **T**. Кожен запис відповідає транзакції, де у відповідному стовпці стоїть 1, якщо елемент присутній в транзакції, і 0 в іншому випадку. Більше того, як видно з таблиці, всі елементи впорядковані в алфавітному порядку (якщо це числа, вони повинні бути впорядковані в числовому порядку).

Виявлення популярних наборів елементів – це операція, яка потребує значних обчислювальних ресурсів і, відповідно, часу. Примітивний підхід до вирішення даного завдання - простий перебір всіх можливих наборів

елементів. Це вимагає порядка $O(2^{|\mathbf{I}|})$ операцій, де $|\mathbf{I}|$ − кількість елементів. Apriori використовує одну з властивостей підтримки, а саме: підтримка будь-якого набору елементів не може перевищувати мінімальної підтримки будь-якої з його підмножин. Причому зворотнє правило не вірне.

Ця властивість носить назву антимонотонності і служить для зменшення розмірності простору пошуку.

Властивості антимонотонності можна дати і інше формулювання: з ростом розміру набору елементів підтримка зменшується, або залишається такою ж. З усього вищесказаного випливає, що будь-який k-елементний набір буде популярним тоді і тільки тоді, коли всі його (k-1)-елементні підмножини будуть популярними.

Всі можливі набори елементів з **I** можна представити у вигляді решітки, що починається з порожньої множини, потім на 1 рівні 1-елементні набори, на 2-му рівні 2-елементні і т.д. На k рівні представлені k-елементні набори, пов'язані з усіма своїми (k-1)-елементнимі підмножиними. Нехай маємо множину транзакцій, яка містить множину предметів **I**={a, b, c, d, e}. Решітка для цієї множини представлена на рис.3.1.

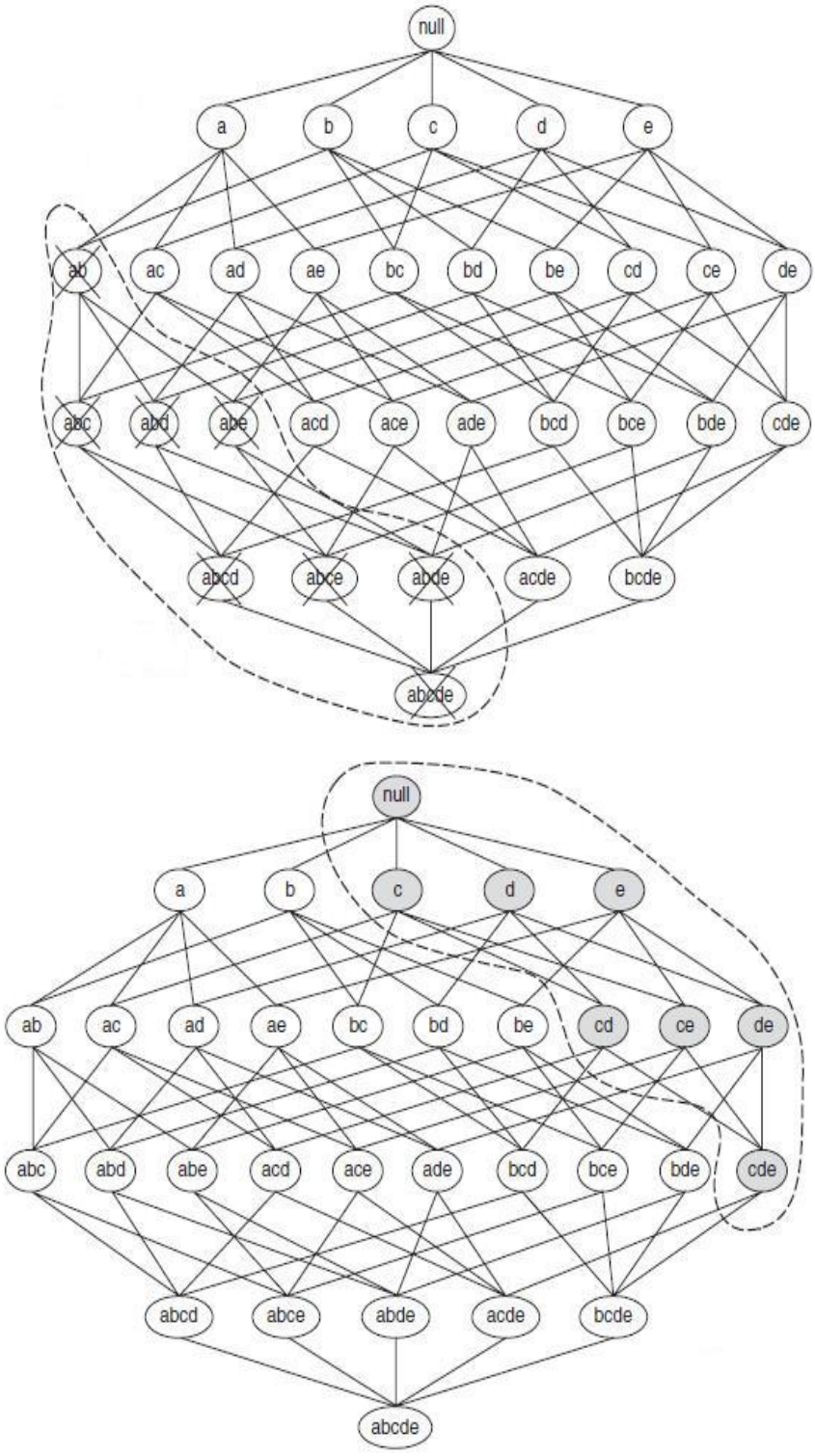


Рис. 3.1 – Решіткова структура для інтерпретації принципу антимонотонності

Принцип антимонотонності формулюється наступним чином:

- якщо кожен елемент множини{c,d,e} − популярний (supp{c}>minsupp, supp{d}>minsupp, supp{e}>minsupp), тоді всі породжені множини популярні;

- якщо кожен елемент множини {a,b} – непопулярний (supp{a}<minsupp, supp{b}<minsupp), тоді всі її породжені підмножини непопулярні.

Застосування принципу антимонотонності в алгоритмі Apriori дозволяє значно зменшити час пошуку асоціативних правил. Блок-схема алгоритму Apriori в частині пошуку популярних наборів наведена на рис. 3.2.

Вхідними величинами для роботи алгоритму є мінімальна підтримка - minsupp, мінімальна достовірність − minconf, N − кількість Top N правил.

Алгоритм рішення задачі включає наступні кроки:

Крок 1. В множині популярних наборів знаходять правила, котрі задовольняють умові:

$$\mathrm{Supp}(\mathbf{X} \Rightarrow \mathbf{Y}) > \mathrm{minsup} \ . \qquad (3.10)$$

Крок 2. Із отриманої множини виключають правила за наступною умовою:

$$\mathrm{conf}(\mathbf{X} \Rightarrow \mathbf{Y}) < \mathrm{minconf} \ . \qquad (3.11)$$

Крок 3. В отриманій множині залишаємо правила, котрі задовольняють умові:

$$\mathrm{Impr}(\mathbf{X} \Rightarrow \mathbf{Y}) > 1 \ . \qquad (3.12)$$

Крок 4. Із цієї множини вибираємо top – N правил із максимальним значенням Leveradge $(\mathbf{X} \Rightarrow \mathbf{Y})$.

В дисертаційній роботі розроблений метод, який дозволяє ітеративно збільшувати значення мінімальної підтримки при заданому значенні мінімальної достовірності і заданому діапазоні кількості асоціативних правил.

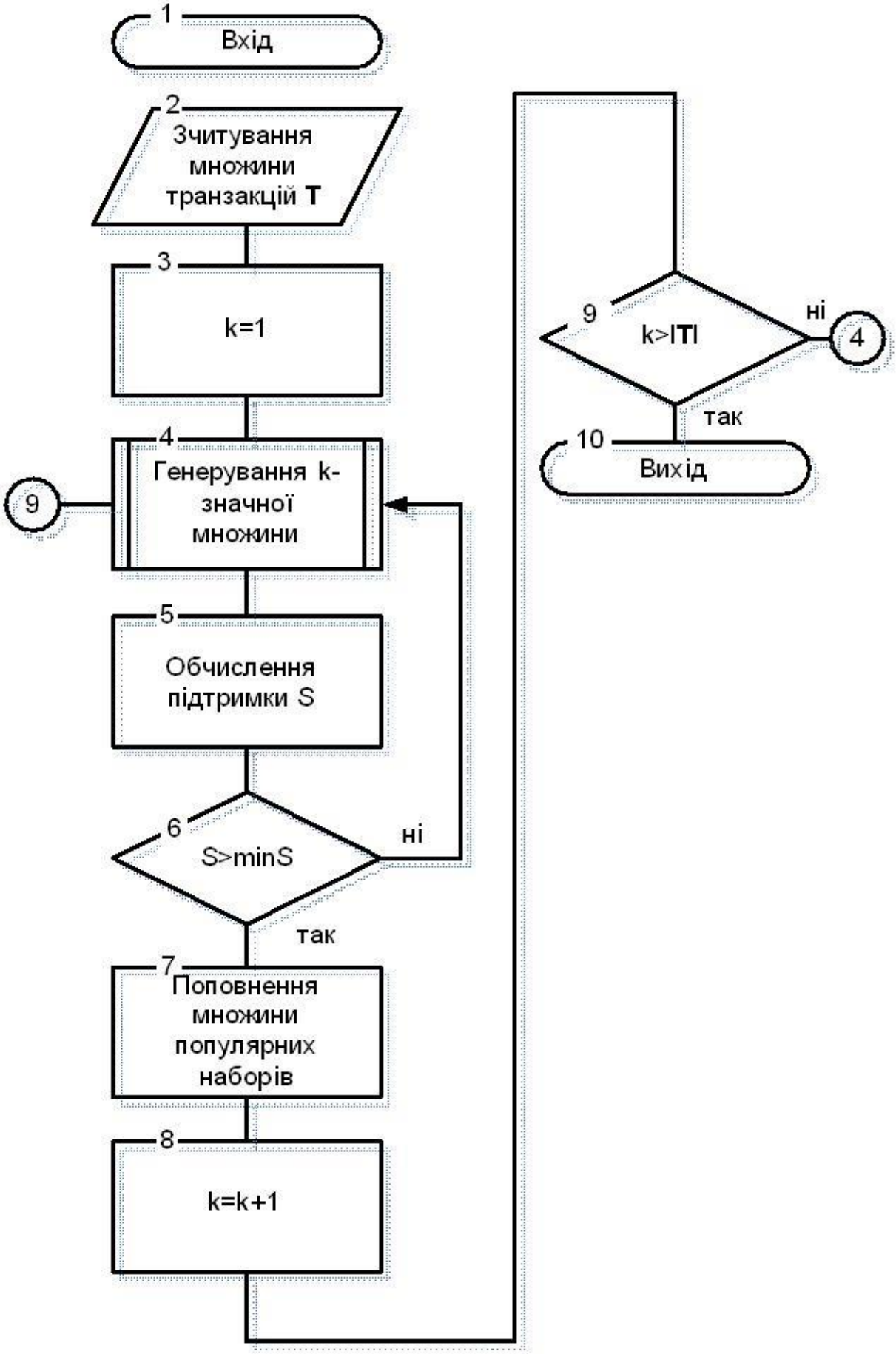


Рис. 3.2 – Блок-схема алгоритму пошуку популярних наборів

Метод полягає в наступному. Нехай задана величина мінімальної підтримки (minsupp), мінімальної достовірності (minconf), мінімальна і максимальна кількість асоціативних правил (N_{min} , N_{max}), потрібно знайти десять правил з найкращими значеннями поліпшення і підсилення. Для початкового значення minsupp знаходяться значення асоціативних правил за допомогою алгоритму Apriori. Якщо кількість знайдених асоціативних правил $N > N_{max}$, тоді значення minsupp збільшується на величину приросту Δminsupp до тих пір, поки $N \in [N_{min}, N_{max}]$. Якщо $N \in [N_{min}, N_{max}]$, тоді метод вибирає $N_1 = 10$ асоціативних правил з найбільшими значеннями Impr і Leverage. В подальшому minsupp збільшується на величину Δminsupp і знову за допомогою алгоритму Apriori метод шукає $N_2 = 10$ асоціативних правил з найбільшими значеннями Impr і Leverage. Серед знайденої множини $\mathbf{N_1} \cup \mathbf{N_2}$ метод вибирає 10 правил з найбільшими значеннями Impr і Leverage. Ітеративний процес триває до тих пір, поки $N < N_{min}$. З отриманої множини $\mathbf{N} = \mathbf{N_1} \cup \mathbf{N_2} \cup \ldots \cup \mathbf{N_i}$ користувач або рекомендаційна системи вибирають необхідну кількість асоціативних правил з максимальними значеннями Impr і Leverage.

При застосуванні методу пошуку асоціативних правил для прогнозування рекомендацій в рекомендаційній системі колаборативної фільтрації матриця користувач-предмет перетворюється в матрицю користувач-інтерес. Не всі вибрані користувачами предмети представляють одинаковий інтерес для користувачів. Рейтинги предметів, які виставляють користувачі, належать інтервалу [0,5] або [0,10]. Перетворення матриці користувач-предмет відбувається за наступним алгоритмом:

1) якщо $r_{ij} < 3$ для інтервалу $r_{ij} \in [0,5]$ або $r_{ij} < 6$ для інтервалу $r_{ij} \in [0,10]$, тоді $r_{ij} = 0$;

2) якщо $r_{ij} \geq 3$ для інтервалу $r_{ij} \in [0,5]$ або $r_{ij} \geq 6$ для інтервалу $r_{ij} \in [0,10]$, тоді $r_{ij} = 1$.

Матриця користувач-предмет перетворюється в матрицю користувач-інтерес, яка містить лише двійкові елементи 0 або 1.

Здійснюється пошук асоціативних правил згідно елементів матриці користувач-інтерес:

$$\mathbf{X} \Rightarrow \mathbf{Y}, \tag{3.13}$$

де множини **X** і **Y** володіють наступними властивостями:

$$\mathbf{X} \subset \mathbf{I}, \mathbf{Y} \subset \mathbf{I}, \mathbf{X} \cap \mathbf{Y} = \varnothing, \tag{3.14}$$

де $\mathbf{I} = \{\mathbf{I_1}, \mathbf{I_2}, \ldots \mathbf{I_n}\}$ − множина предметів, які може рекомендувати користувачам рекомендаційна система.

Кожен рядок матриці користувач-інтерес вважається окремою транзакцією.

Здійснюється пошук асоціативних правил (3.13) для кожного відмінного від нуля значення вектора профілю активного користувача. Це означає, що кардинальне число множини **X** у виразі (3.13) дорівнює одиниці |**X**|=1.

Нехай ми маємо наступну матрицю користувач-інтерес (Рис. 3.3)

	$\mathbf{I_1}$	$\mathbf{I_2}$	$\mathbf{I_3}$	$\mathbf{I_4}$
$\mathbf{U_1}$	0	1	1	0
$\mathbf{U_2}$	1	0	1	1
$\mathbf{U_3}$	1	0	0	1
$\mathbf{U_4}$	0	0	1	1

Рис 3.3 – Приклад матриці користувач−інтерес

Множина **I** складається з наступних елементів **I={$\mathbf{I_1}$, $\mathbf{I_2}$, $\mathbf{I_3}$, $\mathbf{I_4}$}**. Активний користувач $\mathbf{U_2}$. Профіль активного користувача $\mathbf{U_2}$ ={1, 0, 1, 1}.

Рекомендаційна система виконає пошук наступних асоціативних правил

$$i_{12} \Rightarrow \tilde{\mathbf{I}}_1, i_{32} \Rightarrow \tilde{\mathbf{I}}_2, i_{42} \Rightarrow \tilde{\mathbf{I}}_3 , \tag{3.15}$$

де $\tilde{\mathbf{I}}_1 \in \mathbf{I}, \tilde{\mathbf{I}}_2 \in \mathbf{I}, \tilde{\mathbf{I}}_3 \in \mathbf{I}$.

На наступному кроці рекомендаційна система формує множину $\hat{\mathbf{I}} = \tilde{\mathbf{I}}_1 \cup \tilde{\mathbf{I}}_2 \cup \tilde{\mathbf{I}}_3$. З цієї множини система рекомендує користувачу десять предметів з найвищим рейтингом.

3.2. Метод прогнозування рекомендацій для груп користувачів з врахуванням розрідженості матриці користувач-предмет

В даному розділі наведені результати, які отримані автором в роботах [1,3,5,8,10,12]. Формальна постановка задачі прогнозування рекомендацій для груп користувачів полягає в наступному. Нехай $\mathbf{U} = \{\mathbf{U_1}, \mathbf{U_2}, \ldots, \mathbf{U_n}\}$ − множина векторів профілів користувачів, $\mathbf{G} = \{\mathbf{G_1}, \mathbf{G_2}, \ldots, \mathbf{G_M}\}$ − множина груп користувачів, $\mathbf{G_i} = \{\mathbf{U_{1Gi}}, \mathbf{U_{2Gi}}, \ldots, \mathbf{U_{kGi}}\}$ − множина профілів користувачів для групи $\mathbf{G_i}$. Необхідно здійснити прогноз рекомендацій для груп користувачів $\hat{r}_{Gi} = \text{Predict}(\mathbf{G_i})$. Узагальнена схема роботи методу прогнозування рекомендацій для спільнот користувачів наведена на рис.3.4.

Для пошуку груп подібних користувачів використовуються методи кластеризації [41,57].

Особливістю матриці користувач-предмет є те, що вона містить значну кількість нульових елементів. Кількість ненульвих елементів не перевищує 10% від загальної кількості елементів матриці предмет-користувач [41,57]. Тому для кластеризації користувачів у групи доцільно використовувати демографічні характеристики користувачів. Основними демографічними атрибутами користувачів є наступні: вік, стать, освіта, рід занять. Вік − це числовий атрибут, стать, освіта, рід занять − категоріальні атрибути.

Нехай рейтинговий вектор профілю *i*-того користувача задається наступним вектором (3.16)

$$\mathbf{U_i} = (u_{1i}, u_{2i}, \ldots, u_{mi}) \ , \qquad (3.16)$$

де u_{ji} - рейтингова оцінка *j*- того предмета *i*- тим користувачем.

Розширимо цей вектор за допомогою демографічних атрибутів користувача (3.17)

$$\mathbf{U_i^{ext}} = (u_{1i}, u_{2i}, \ldots, u_{mi}, d_{1i}, d_{2i}, d_{3i}, d_{4i}) \ , \qquad (3.17)$$

де $d_{1i}, d_{2i}, d_{3i}, d_{4i}$ – категоріальні атрибути користувача.

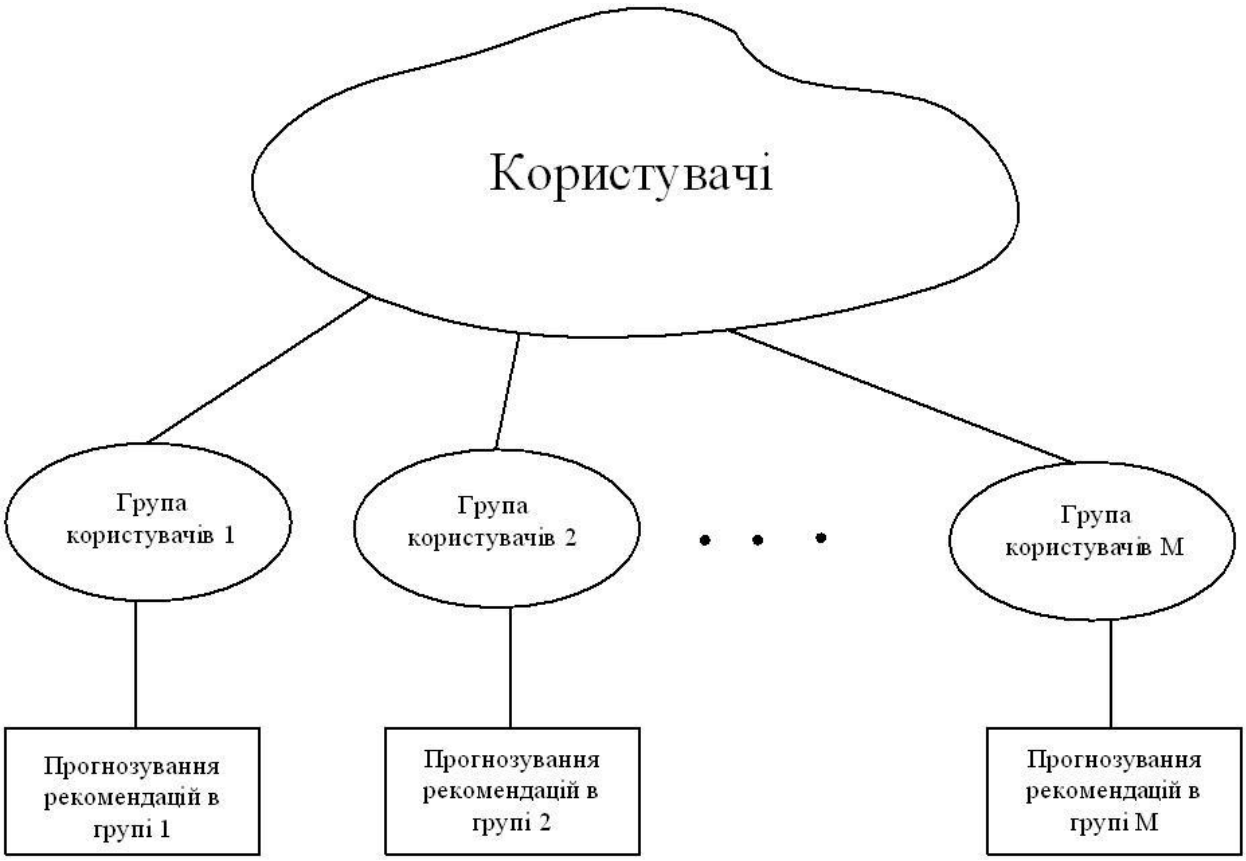


Рис.3.4 – Узагальнена схема роботи методу прогнозування рекомендацій для груп користувачів

Для спрощення опису методу будемо позначати вектор $\mathbf{U_i^{ext}}$ за допомогою вектора $\mathbf{X_i} = (x_{1i}, x_{2i}, \ldots, x_{ni})$.

Категоризація числового атрибута "вік" наведена в табл. 3.3.

Таким чином отримуємо мішаний вектор профідю користувача, який містить числові і категоріальні значення (3.16).

Кластеризація мішаних векторів профілів користувачів здійснюється за допомогою методу мішаної кластеризації, який розроблений в роботі [1]. Метод мішаної кластеризації базується на розрахунку щільності розміщення мішаних векторів профілів користувачів і визначає кількість і положення центрів кластерів. Щільність визначається як кількість векторів профілів користувачів, які знаходяться в околі радіусом d_c біля кожного користувача (3.18), (3.19) :

$$\rho_i = \sum_{j=1}^{N} f(d_{ij} - d_c), \tag{3.18}$$

де d_{ij} – відстань між *i*- тим та *j*- тим векторами профілів користувачів;

d_c – порогове значення;

N – кількість користувачів.

$$f(x) = \begin{cases} 1, x = d_{ij} - d_c \le 0 \\ 0, x = d_{ij} - d_c > 0 \end{cases}. \tag{3.19}$$

Таблиця 3.3 – Категоризація атрибуту користувача - вік

Вік			
вік<18	18<вік<30	30<вік<50	50<вік
1	2	3	4

Відстань між векторами профілів користувачів обчислюється як середня зважена сума відстаней між атрибутами векторів профілів користувачів:

$$d_{ij} = D(\mathbf{X_i}, \mathbf{X_j}) = \frac{\sum_{k=1}^{N} w_{ij}^k l_{ij}^k}{\sum_{k=1}^{N} w_{ij}^k}, \quad (3.20)$$

де $w_{ij}^k = 0$, якщо атрибут d_{ij}^k відсутній і $w_{ij}^k = 1$ в інакшому випадку,

l_{ij}^k - відстань між *i*- тим і *j*- тим атрибутамим об'єктів $\mathbf{X_i}$ і $\mathbf{X_j}$.

Відстань l_{ij} обчислюється окремо для числових і категоріальних атрибутів векторів $\mathbf{X_i}$ і $\mathbf{X_j}$.

Для числових атрибутів:

$$d_{ij}^k = \frac{\left| x_i^k - x_j^k \right|}{\max z_{ij}^k - \min z_{ij}^k}, \quad (3.21)$$

де $\max z_{ij}^k$ – максимальне значення в множині атрибутів векторів $\mathbf{X_i}$ і $\mathbf{X_j}$,

$\min z_{ij}^k$ – мінімальне значення в множині атрибутів векторів $\mathbf{X_i}$ і $\mathbf{X_j}$.

Для категоріальних атрибутів:

$$l_{ij}^k = \begin{cases} 0, x_i^k = x_j^k \\ 1, x_i^k \neq x_j^k \end{cases}. \quad (3.22)$$

Для знаходження положення і кількості центрів кластерів визначаються мінімальні відстані між множинами об'єктів з різними щільностями:

$$\delta_i = \min_j (d_{ij}), \rho_j > \rho_i . \quad (3.23)$$

Будуються два вектори в порядку спадання значень $\rho_a > \rho_b$, $\delta_a > \delta_b$ (рис.3.5)

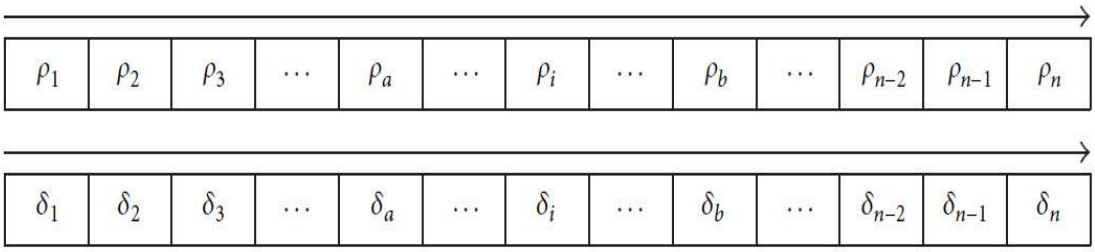
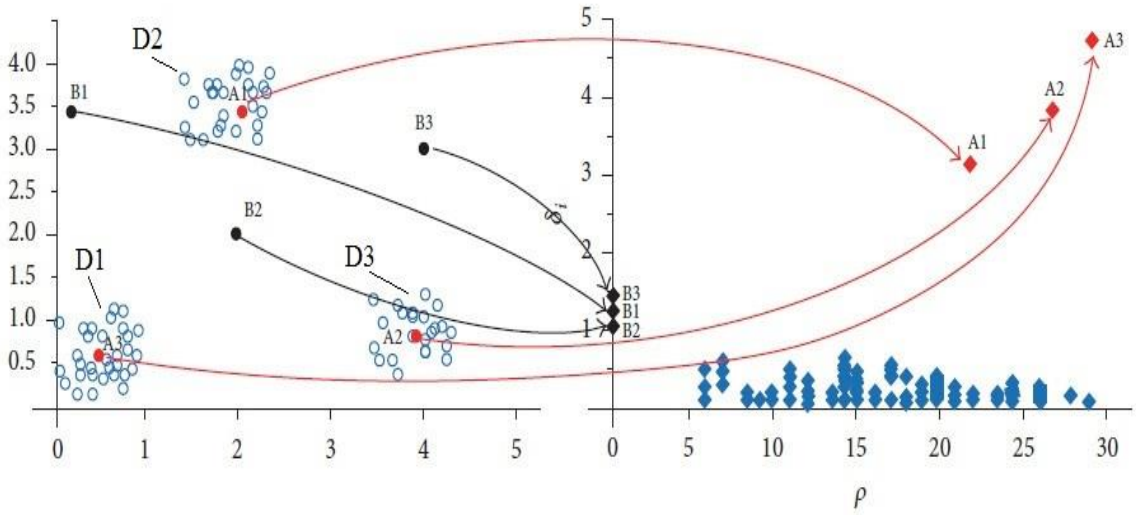


Рис.3.5 – Вектори щільностей і відстаней між множинами

Будемо вважати, що $\rho_1 - \rho_a$ − це "великі" значення щільностей, $\rho_b - \rho_n$ − це "малі" значення щільностей. Подібні умови накладемо на $\delta_1 \div \delta_a$ і $\delta_b \div \delta_n$.

Якщо виконується умова $\rho_i \in (\rho_1, \rho_a)$ і $\delta_i \in (\delta_1, \delta_a)$, тоді $\mathbf{X_i}$ центр наступного кластера.

Якщо виконується умова $\rho_i \in (\rho_b, \rho_n)$ і $\delta_i \in (\delta_1, \delta_a)$, тоді $\mathbf{X_i}$ обєкт "шуму" і в подальших розрахунках не враховується. Графічна інтерпретація методу в 2D координатах наведена на рис. 3.6.



а) б)

Рис. 3.6 – Графічна інтерпретація мішаного методу кластеризації в 2D координатах

На рис. 3.6 а) представлений приклад розподілу точок в 2D просторі.

На рис. 3.6 б) представлений приклад розподілу параметрів ρ і δ для попереднього розподілу. D1, D2, D3 − області згущення точок з великою

щільністю. A1, A2, A3 − центри кластерів, B1, B2, B3 − точки "шуму". Після визначення центрів кластерів здійснюється поділ об'єктів на кластери. Для пошуку кластерів в системі прогнозування рекомендацій передбачено модифікований метод сканування щільності і модифікований метод k - середніх.

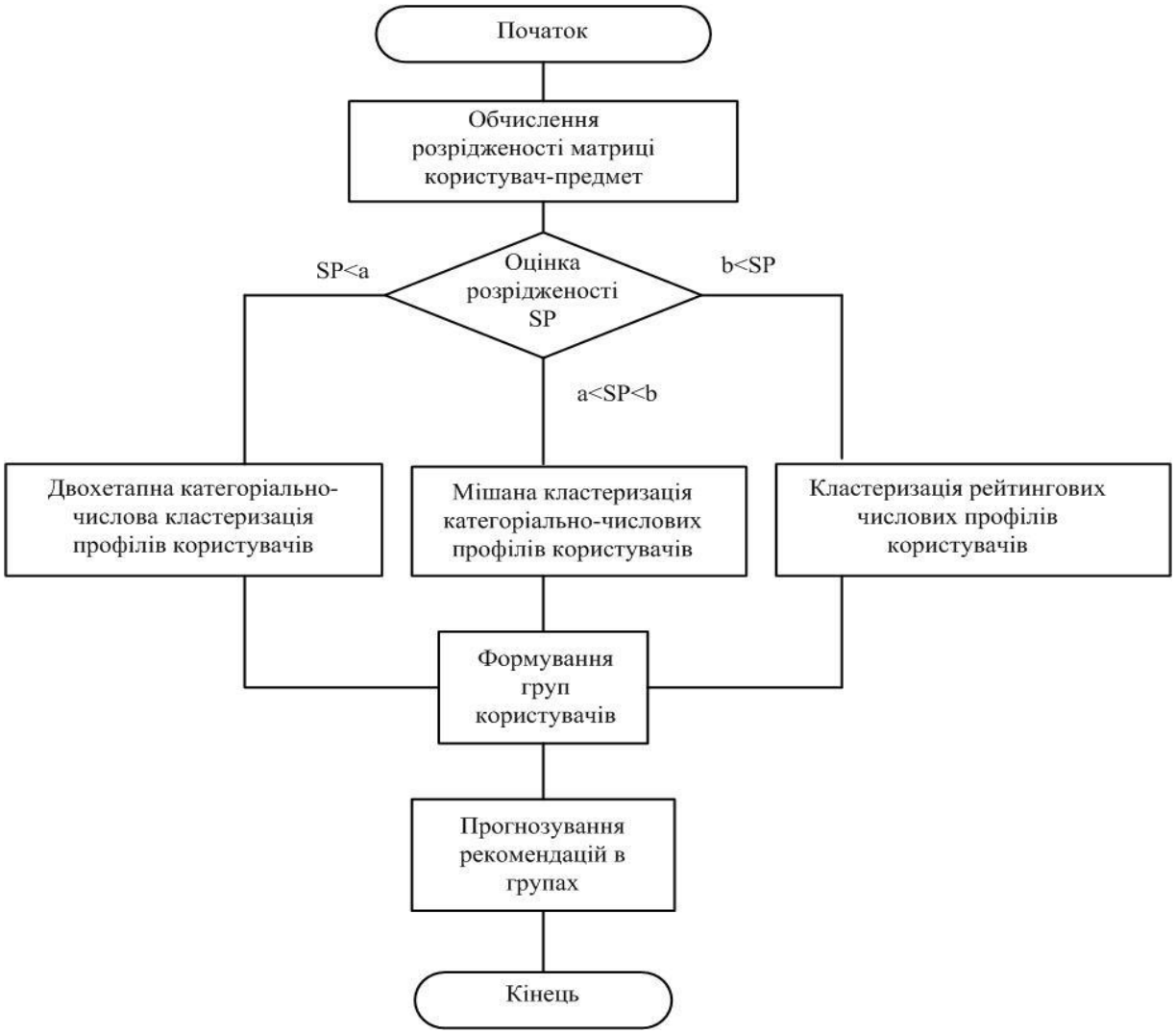


Рис 3.7 – Структурна схема гібридного методу кластеризації для пошуку груп подібних користувачів

Результати пошуку груп користувачів в значній мірі залежать від розрідженості матриці користувач-предмет. Розрідженість матриці користувач-предмет може бути розрахована за допомогою наступного виразу:

$$SP = 1 - \frac{nR}{nUSER * nITEM}, \quad (3.24)$$

де nR − кількість відмінних від нуля елементів матриці користувач-предмет;

nUSER − кількість користувачів системи;

nITEM − кількість предметів в системі.

Розрідженість матриці користувач-предмет використовується в гібридному методі пошуку груп користувачів. Узагальнена структурна схема двоетапного гібридного методу категоріально-числової кластеризації представлена на рис. 3.8.

Рис.3.8 – Узагальнена схема двоетапного методу категоріально-числової кластеризації

Гібридний метод кластеризації для пошуку груп користувачів включає наступні методи: модифікований метод неієрархічної кластеризації числових векторів профілів користувачів, який базується на методі k - середніх; метод мішаної кластеризації категоріально-числових векторів профілів користувачів;

двоетапний метод категоріально-числової кластеризації. Вибір методу здійснюється за допомогою оцінки розрідженості матриці користувач-предмет і двох параметрів a і b. При малій розрідженості використовується модифікований метод неієрархічної кластеризації числових векторів профілів користувачів, який базується на методі k - середніх. При середньому значенні розрідженості використовується метод мішаної кластеризації категоріально-числових векторів профілів користувачів. При великій розрідженості використовується двоетапний метод категоріально-числової кластеризації.

На першому етапі здійснюється категоріальна кластеризація векторів демографічних профілів користувачів і будуються групи користувачів, які близькі за своїми демографічними характеристиками. На другому етапі здійснюється числова кластеризація векторів профілів користувачів, які містять числові рейтингові оцінки предметів. Прогнозування рекомендацій в отриманих групах може бути виконано як класичним методом колаборативного прогнозування користувач-користувач або предмет-предмет, так і методом прогнозування в групі. Для категоріальної кластеризації векторів демографічних профілів користувачів використовується модифікований метод ROCK (A Robust Clustering Algorithm for Categorical Attributes) [110].

ROCK являє собою агломеративний ієрархічний алгоритм кластеризації категоріальних атрибутів. Замість звичної міри близькості для методів числової кластеризації (норма Евкліда) в методі ROCK вводиться міра зв'язу між атрибутами і множинами атрибутів, яка не задовольняє аксіомам метричного простору, однак ефективна при кластеризації категоріальних векторів в n- мірному просторі. Однак метод ROCK може будувати хибні кластери на кінцевій стадії кластеризації. Модифікований метод ROCK вимагає менше часу для розрахунку і вирішує проблему кластеризації на кінцевій стадії.

3.3. Моделі прогнозування рекомендацій для предметів у методі прогнозування рекомендацій для груп користувачів

В гібридному методі прогнозування рекомендацій відбувається прогнозування рекомендацій для груп користувачів. Тому виникає задача розроблення моделей прогнозування рекомендацій для груп користувачів. За основу побудови таких моделей приймаються індивідуальні моделі користувачів, які визначають інтерес кожного користувача до множини предметів рекомендаційної системи. Індивідуальна модель користувача – це вектор профілю користувача, який визначає інтерес користувача до кожного предмета. Групову модель можна розглядати як синтез моделей користувачів, побудований комбінуванням векторів профілів користувачів, які входять в групу. Побудова групової моделі базується на принципах теорії колективного вибору [111,112]. Модель вибору для групи враховує інтереси користувачів, які входять в групу. До таких моделей належать: адитивна утилітарна модель, мультиплікативна утилітарна модель, модель голосування схваленням, модель найменшого задоволення, модель найбільшого задоволення, модель середнього значення без найменшого задоволення.

Адитивна утилітарна модель

При цій стратегії обчислюється сума рейтингів для кожного предмета в таблиці користувач-предмет.

Таблиця 3.4 – Матриця користувач-предмет для адитивної утилітарної моделі

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
u_1	8	10	7	10	9	8	10	6	3	6
u_2	7	10	6	9	8	10	9	4	4	7
u_3	5	1	8	6	9	10	3	5	7	10
Групова оцінка	20	21	21	25	26	28	22	15	14	23

В результаті формується вектор адитивної корисності предметів в групі табл. 3.4.

Список рекомендованих предметів формується по спаданню адитивної оцінки.

Мультиплікативна утилітарна модель

Таблиця 3.5 – Матриця користувач-предмет для мультиплікативної утилітарної моделі

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
u_1	8	10	7	10	9	8	10	6	3	6
u_2	7	10	6	9	8	10	9	4	4	7
u_3	5	1	8	6	9	10	3	5	7	10
Групова оцінка	280	100	336	540	648	800	270	120	84	420

При цій стратегії обчислюється добуток рейтингів для кожного предмета в таблиці користувач-предмет (табл. 3.5). В результаті формується вектор мультиплікативної корисності предметів в групі.

Модель голосування схваленням

При цій стратегії користувач «голосує» за певні предмети. Вважається, що користувач проголосував за певний предмет, якщо в комірці матриці користувач – предмет стоїть 1 і не проголосував, якщо стоїть 0. Припустимо, що кожен користувач голосує за предмет, який має рейтингове значення вище за певне порогове значення (наприклад 5). Тоді матриця користувач-предмет (табл. 3.6) перетворюється в наступну матрицю:

Таблиця 3.6 – Матриця користувач-предмет для моделі голосування схваленням

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
u_1	1	1	1	1	1	1	1	1	0	1
u_2	1	1	1	1	1	1	1	0	0	1
u_3	0	0	1	1	1	1	0	0	1	1
Групова оцінка	2	2	3	3	3	3	2	1	1	3

Модель найменшого задоволення

В рейтинг групової оцінки входить найнижча оцінка, яку виставили користувачі кожному предмету (табл. 3.7).

Таблиця 3.7 – Матриця користувач-предмет для моделі найменшої оцінки задоволення

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
u_1	8	10	7	10	9	8	10	6	3	6
u_2	7	10	6	9	8	10	9	4	4	7
u_3	5	1	8	6	9	10	3	5	7	10
Групова оцінка	5	1	6	6	8	8	3	4	3	6

Переважно така модель використовується для малих груп користувачів. Недолік такої моделі проявляється тоді, коли більшість користувачів в системі оцінили предмет високими оцінками і лише один користувач виставив низьку оцінку.

Модель найбільшого задоволення

В рейтинг групової оцінки входить найвища оцінка, яку виставили користувачі кожному предмету (табл. 3.8).

Таблиця 3.8 – Матриця користувач-предмет для моделі з найбільшою оцінкою задоволення

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
u_1	8	10	7	10	9	8	10	6	3	6
u_2	7	10	6	9	8	10	9	4	4	7
u_3	5	1	8	6	9	10	3	5	7	10
Групова оцінка	8	10	8	10	9	10	10	6	7	10

Модель середнього значення без найменшого задоволення

Для кожного предмета оцінка вибирається як сума оцінок, які більші за певну порогову величину. Предмети, які отримали хоча б одну оцінку меншу порогового значення, не включаються в групову оцінку. Матриця користувач-предмет, яка наведена в табл. 3.9, побудована для порогового значення 4:

Таблиця 3.9 – Матриця користувач-предмет для моделі середнього значення без найменшого задоволення

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
u_1	8	10	7	10	9	8	10	6	3	6
u_2	7	10	6	9	8	10	9	4	4	7
u_3	5	1	8	6	9	10	3	5	7	10
Групова оцінка	20	–	21	25	26	28	–	15	–	23

Результати оцінки наведених вище моделей необхідно привести до інтервалу [0,5] або інтервалу [0,10]. Моделі найменшого задоволення і найбільшого задоволення дають результати, які належать вказаним інтервалам і не вимагають приведення. Для адитивної утилітарної моделі, мультиплікативної утилітарної моделі, моделі середнього значення без найменшого задоволення використовується наступний вираз:

$$ntw\_group\_rating = \frac{group\_rating * \max\_rating}{\max\_group\_rating}, \qquad (3.25)$$

де $group\_rating$ − рейтинг, який створений вибраною стратегією моделювання;

$\max\_group\_rating$ − максимальний рейтинг, який користувач може виставити предмету;

$\max\_rating$ − максимальне значення рейтингу, який може бути створений стратегією моделювання.

В моделі голосування схваленням вибір предметів здійснюється із

вектора групової оцінки, в якому предмети впорядковані по спаданню значення сумарної групової оцінки.

3.4. Формальні теоретико-множинні моделі прогнозування рекомендацій для перехресних продаж (cross-selling) і додаткових продаж (up-selling)

Cross-selling і Up-selling (перехресні продажі і підняття суми продажу) це два популярних підходи до підвищення прибутку інтернет-магазину.

Up-selling або підняття суми продажу - це мотивація покупця витратити більше грошей у інтернет-магазині, наприклад, купити дорожчу модель того ж продукту, додати опції або послуги до продукту, який купує користувач.

Cross-selling або перехресні продажі - це також мотивація покупця витратити більше грошей, але вже через продаж товарів з інших категорій, ніж спочатку обрана користувачем, тобто в першу чергу продаж супутніх товарів.

Метою перехресних продажів може бути або збільшення доходу, отримуваного від клієнта, або захист відносин з клієнтом або клієнтами. На відміну від придбання нового бізнесу, крос-продажі пов'язані з ризиком, що існуючі відносини з клієнтом можуть бути порушені. З цієї причини важливо забезпечити, щоб додатковий продукт або послуга, що продаються клієнту або клієнтам, збільшували цінність, яку клієнт або клієнти отримують від інтернет-магазину.

Підняття суми продажу - це практика, при якій інтернет-магазин намагається переконати клієнтів придбати дорожчий предмет, апгрейд або додатковий предмет для збільшення прибутку з продажу. Наприклад, інтернет-магазин може запропонувати користувачу купити нову дорожчу модель товару, а не більш дешеву поточну модель, вказавши на додаткові функції нового товару.

Позначимо через **U** множину профілів користувачів

$$\mathbf{U} = (\mathbf{U_1}, \mathbf{U_2}, \ldots, \mathbf{U_m}), \tag{3.26}$$

де $\mathbf{U_i} = (u_{1i}, u_{2i}, \ldots, u_{ni})$ – профіль користувача $\mathbf{U_i}$.

Позначимо через **I** множину профілів предметів

$$\mathbf{I} = (\mathbf{I_1}, \mathbf{I_2}, \ldots, \mathbf{I_n}), \tag{3.27}$$

де $\mathbf{I_j} = (i_{1j}, i_{2j}, \ldots, i_{mj})$ – профіль предмета $\mathbf{I_j}$.

Позначимо через $\mathbf{G_I}$ множину категорій предметів

$$\mathbf{G_I} = (\mathbf{G_{1I}}, \mathbf{G_{2I}}, \ldots, \mathbf{G_{kI}}), \tag{3.28}$$

де $\mathbf{G_{lI}} = (\tilde{i}_1, \tilde{i}_2, \ldots, \tilde{i}_p)$ – множина предметів, які належать до категорії $\mathbf{G_{lI}}$.

Нехай користувач u_{ij} вибирає предмет из категорії $\mathbf{G_{lI}}$. Тоді підхід перехресних продаж пропонує користувачу предмети из інших категорій:

$$\mathbf{G_{cross}} = \{\mathbf{G}_{jI}\}, i \neq l \ . \tag{3.29}$$

Режим додаткових продаж пропонує предмети з тієї ж самої категорії $\mathbf{G_{lI}}$. При цьому для додаткових продаж пропонуються предмети, для яких асоціативний аналіз показав, що вони вибрані разом з предметами, які входять в профіль користувача $u_{ij} = (\hat{i}_1, \hat{i}_2, \ldots, \hat{i}_r)$:

$$\hat{i}_j \Rightarrow \hat{\mathbf{I}}_i, \ j = 1, 2, \ldots r, \tag{3.30}$$

де $-\ \hat{\mathbf{I}}_i \in \mathbf{G_{lI}}, i = 1, 2, \ldots, r$.

Тоді предмети для додаткових продаж вибираються із множини $\mathbf{G_{UP}} = \mathbf{I}_i \cup \mathbf{I}_j$, $i, j = 1, 2, \ldots, p, i \neq j$.

3.5. Метод збільшення різноманітності прогнозованих предметів

В даному розділі наведені результати, отримані автором в роботі [13]. До основних задач рекомендаційних систем належать (рис. 3.9):

1. Конверсія − відвідувач інтернет-магазину стає покупцем;
2. Лояльність − користувач неодноразово здійснює покупки в інтернет – магазині;
3. Забезпечення продаж супутніх предметів;

4. Вирішення проблеми « довгого хвоста » (long tail).

Рис.3.9 – Основні задачі рекомендаційних систем в галузі електронної комерції

Переважно рекомендаційні системи пропонують предмет із великою популярністю. Кількість таких предметів не перевищує 20% від загальної кількості предметів, які маже запропонувати інтернет-магазин. Завдання рекомендаційної системи полягає в тому, щоби для користувача стали доступними предмети із області «довгого хвоста».

У даній роботі пропонується метод і алгоритм збільшення різноманітності предметів в методі колаборативної фільтрації. Нехай існує множина предметів $\mathbf{S} = \{s_1, s_2, \ldots, s_n\}$, ця множина має високий рівень різноманітності, якщо існує велика відмінність між предметами в множині. Нехай відмінність між предметом S_i і S_j обчислюється за виразом

$dist(s_i, s_j)$. Тоді різноманітність множини предметів в цілому може бути розрахована за формулою:

$$diversity(\mathbf{S}) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} dist(s_i, s_j), s_i \in \mathbf{S}, s_j \in \mathbf{S} \ . \quad (3.31)$$

Рекомендаційна система пропонує користувачу Top n предметів. Переважно це десять предметів з найвищим прогнозованим рейтингом. Однак серед цих предметів можуть бути такі, які майже не відрізняються за своїми характеристиками. Матриця предмет користувач в методі колаборативних фільтрації характеризується великою розрідженістю. Кількість ненульових елементів не перевищує 6-7%. Для прогнозування рейтингів предметів використовується метод зваженої суми. Коефіценти подібності для векторів профілів користувачів враховують тільки співпадаючі значення рейтингових оцінок. Це призводить до похибок в прогнозуванні рекомендацій. Традиційно рекомендаційні системи не враховують різноманітність при прогнозуванні рекомендацій. Структурна схема процесу прогнозування рекомендацій з урахуванням різноманітності представлена на рис. 3.10.

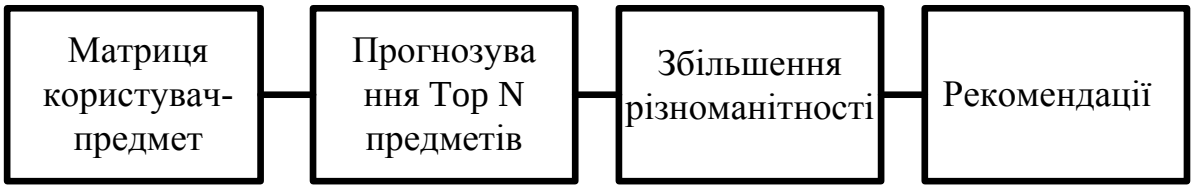


Рис.3.10 – Структурна схема процесу прогнозування рекомендацій із збільшенням різноманітності

Для розрахунку різноманітності використовується такий вираз:

$$FDIV(i, \mathbf{U}, \mathbf{R}) = \alpha F_1(i) + (1-\alpha) F_2(i, \mathbf{R}) * (\beta F_3(i, \mathbf{U}) + (1-\beta) F_4(i, \mathbf{U})) \ , \quad (3.32)$$

де i – предмет, різноманітність якого досліджується;

$\mathbf{U}$ – профіль активного користувача;

$\mathbf{R}$ – множина предметів, які зпрогнозовані рекомендаційною системою;

$\alpha, \beta \in (0,1]$ – керуючі параметри.

Компоненти виразу (3.32) мають наступне призначення:

$F_1(i)$ – оцінка пріоритетності прогнозованого рейтингу для *i*-того предмета;

$F_2(i, \mathbf{R})$ – оцінка відмінності і-того предмета від решти предметів з множини $\mathbf{R}$;

$F_3(i, \mathbf{U})$ – оцінка переваг і-того предмета для активного користувача;

$F_4(i, \mathbf{U})$ – оцінка новизни предмету і по відношенню до множини $\mathbf{R}$.

Для обислення компонент формули (3.31) використовуються наступні вирази

$$F_1(i) = \frac{pr(i)}{MaxRating} \ , \tag{3.33}$$

де $pr(i)$ − прогнозоване значення рейтингу для і-того предмета;

$MaxRating$ − максимальне можливе значення рейтингу;

$$F_2(i) = \frac{1}{|\mathbf{R}|} \sum_{r \in \mathbf{R}} diss(i, r) \ , \tag{3.34}$$

де $diss(i,r)$ − оцінка відмінності між і-тим і r-тим предметами;

$$F_3(i, \mathbf{U}) = \frac{\sum_{u \in \mathbf{U}} sim(i, u)\, r(u, \mathbf{U})}{\sum_{u \in \mathbf{U}} r(u, \mathbf{U})} \ , \tag{3.35}$$

де $sim(i,u)$ − подібність між *i*-тим і *u*-тим предметами;

$r(u, \mathbf{U})$ − рейтинг предмета u в профілі активного користувача $\mathbf{U}$;

$$F_4(i, \mathbf{U}) = div(i, \{\mathbf{N} \cup \mathbf{U}\}) \ , \tag{3.36}$$

де $\mathbf{N}$ − множина об’єктів з околу активного користувача $\mathbf{U}$.

Нехай множина рекоменованих предметів **R**. Для збільшення різноманітності ця множина предметів ітеративно збільшується за рахунок предметів з множини **N**. Множина **N** утворюється, як множина предметів з околу предметів для активного користувача.

Алгоритм збільшення різноманітності предметів наведено на рис. 3.11.

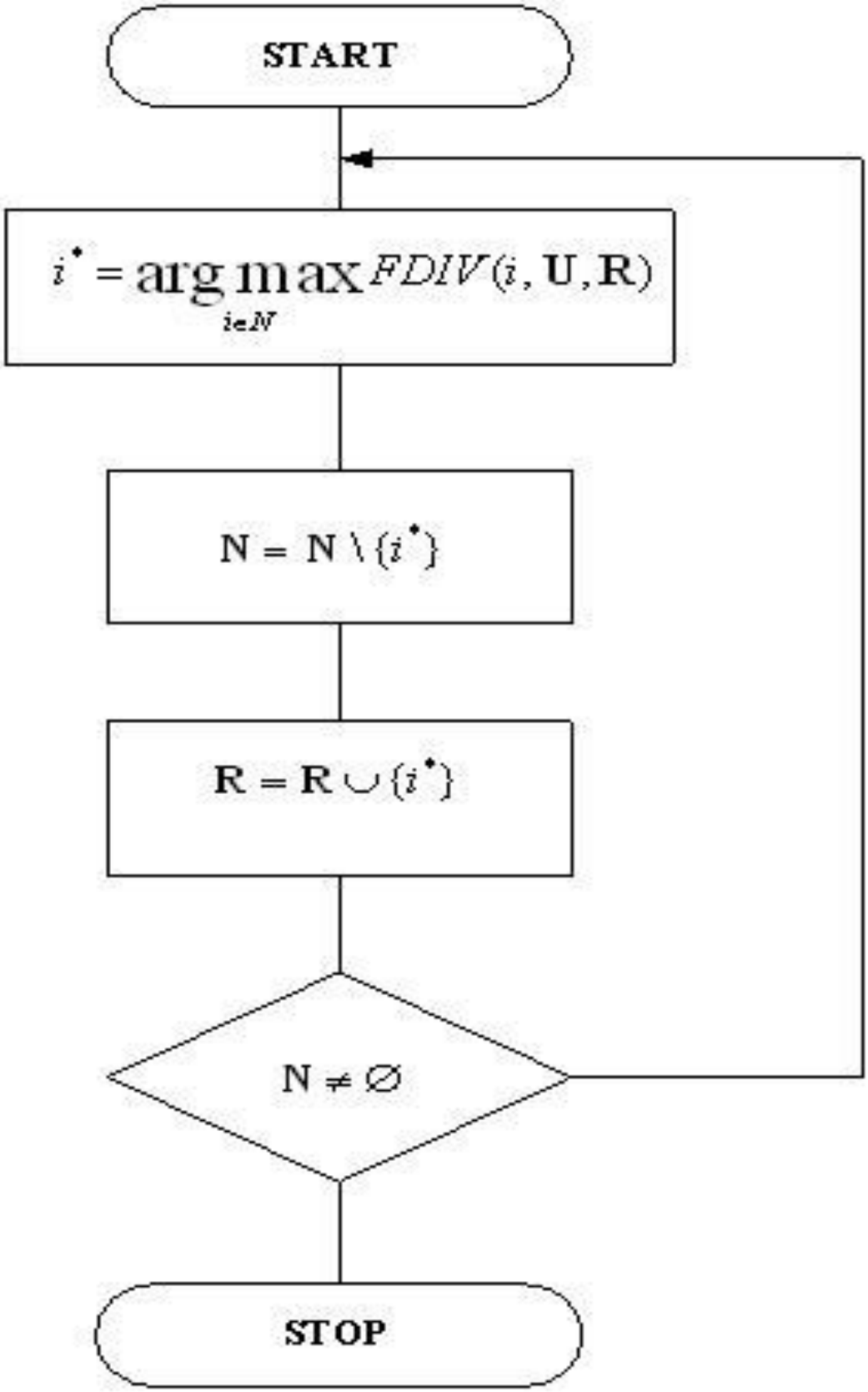


Рис.3.11 – Блок-схема алгоритму збільшення різноманітності предметів.

3.6. Висновки до Розділу 3

Новим і перспективним напрямком дослідження в області рекомендаційних систем є розроблення гібридних методів прогнозування рекомендацій. Такі методи базуються на методі колаборативної фільтрації і моделях інтелектуального аналізу даних. У **третьому розділі** дисертаційної роботи автором розроблено метод прогнозування рекомендацій на основі алгоритму пошуку асоціативних правил Apriori для зареєстрованих в системі користувачів. Метод враховує рівень інтересу користувачів. Пошук асоціативних правил здійснюється на основі матриці користувач-предмет, в якій кожна рейтингова оцінка, яка більша за наперед задану порогову величину, приймається рівною одиниці. Всі решта елементи матриці користувач-предмет приймаються рівними нулю. Кожен вектор профілю користувача вважається окремою транзакцією. Асоціативні правила будуються на множині векторів профілів користувачів, як на множині транзакцій. В подальшому асоціативні правила безпосередньо використовуються для прогнозування рекомендацій і не вимагають застосування коефіцієнтів подібності векторів профілів і методу зваженої суми. Розроблено гібридний метод зменшення розмірності простору векторів профілів користувачів і векторів профілів предметів, який базується на виділенні груп користувачів з подібними інтересами. Метод адаптується до розрідженості матриці користувач-предмет і включає чітку кластеризацію, мішану кластеризацію, категоріальну кластеризацію. Розроблено метод мішаної кластеризації, який не вимагає початкового задання кількості і положення центрів кластерів. Розроблено метод збільшення різноманітності прогнозованих предметів, який дозволяє розв’язувати проблему «довгого хвоста».

Основні положення цього розділу викладені у публікаціях автора [1, 3-5, 8, 10, 12-14].

РОЗДІЛ 4. РОЗРОБЛЕННЯ І ДОСЛІДЖЕННЯ МАТЕМАТИЧНОГО І ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ РЕКОМЕНДАЦІЙНОЇ СИСТЕМИ ІНТЕРНЕТ-МАГАЗИНУ

Під час виконання дисертаційної роботи було розроблено рекомендаційну систему для дослідження роботи та оцінки результатів алгоритмів рекомендацій, що застосовуються в рекомендаційних системах інтернет-магазинів. Систему можна розширювати, що надає можливість модифікації існуючих алгоритмів та створення нових, а також комбінації декількох для отримання кращого результату.

У системі реалізовано алгоритми пошуку схожості користувачів за:

1. Косинусною мірою подібності;
2. Евклідовою мірою подібності;
3. Манхетенівською мірою подібності;
4. Коефіцієнтом кореляції Пірсона;
5. Коефіцієнтом подібності на основі оберненої евклідової відстані;
6. Коефіцієнтом подібності, який враховує демографічні характеристики користувачів.

Вхідні дані для системи зберігаються у вигляді БД MS SQL 2008. Це надає можливість використовувати розроблену систему як частину більшого програмного комплексу.

Розроблена система має на меті забезпечити:

- зручний інтерфейс користувача з текстовим (табличним) відображенням результатів і проміжних кроків надання рекомендацій;
- вибір алгоритму схожості, для отримання різноманітних рекомендацій і оцінки ефективності алгоритму, в тому числі змінюючи кількість даних для порівняння;

Теоретична та практична цінність роботи полягає у наступному:

- дослідження проблеми ефективності рекомендаційних алгоритмів;
- аналіз та вдосконалення існуючих, розроблення нових алгоритмів рекомендації, а також дослідження використання комбінацій алгоритмів, які дають кращі результати;

4.1. Вибір засобів розроблення системи

Для забезпечення можливості розширення системи було вибрано об’єктно-орієнтований підхід.

Систему було розроблено у середовищі MS Visual Studio 2010. Середовище було вибрано з огляду на його зручність побудови користувацьких інтерфейсів. Мова розробки – C#, вибрана через високу зручність і гнучкість виконання кодів.

Для роботи з самою БД використовувалась MS SQL Management Studio 2008. Вибір зумовлений тим, що це активне середовище для роботи з СУБД MS SQL 2008.

Для доступу до БД використовується ORM технологія EntityFramework розроблена корпорацією Microsoft. Вона вибрана за високу швидкість і зручність роботи з нею.

4.2. Структура рекомендаційної системи

Система реалізована у вигляді одного виконуваного файлу, який виконує усі основні функції.

Структурна схема рекомендаційної системи представлена на рис. 4.1.

Підсистема керування діалогом дозволяє вводити керуючі дані для ініціалізації користувача в рекомендаційній системі, вибору моделі, методу прогнозування, метрики подібності, відображає результати прогнозування рекомендацій. Модуль входу в рекомендаційну систему ініціалізує користувача в системі.

Модуль вибору моделі прогнозування дозволяє вибрати модель користувач-користувач або предмет-предмет. Модуль вибору методу прогнозування дозволяє вибрати один із методів колаборативної фільтрації або один із гібридних методів прогнозування рекомендацій.

Модуль вибору метрики подібності дозволяє вибрати один із методів розрахунку метрики подібності.

Модуль прогнозування рекомендацій здійснює прогнозування рекомендацій за допомогою вибраної моделі, методу, метрики і використовує в якості вхідних даних матрицю користувач-предмет із бази даних профілів користувачів і предметів.

Модуль оцінки точності прогнозування здійснює поділ матриці користувач-предмет на тестову і прогнозовану частини і обчислює точність прогнозування.

Модуль візуалізації результатів за допомогою підсистеми керування діалогом відображає результати прогнозування рекомендацій.

Технічні засоби, які використовувались для розроблення рекомендаційної системи:

Апаратні:

- Процесор AMD FX-6350 3.9GHz
- ОЗП 4 ГБ
- HDD 1500 ГБ
- Відеоадаптер NvidiaGeForceGTX 550

Операційна система: Windows 7 Ultimate

Середовище розробки: MS Visual Studio 2010,
MS SQL Management Studio 2008.

Програма призначена для аналізу даних алгоритмами, які в ній реалізовані та для надання рекомендацій на основі схожості користувачів і фільмів.

Система є сумісна з версіями Windows починаючи з Windows XP Home Edition з встановленим **.NET Framework 4**. Вхідні дані зберігаються в БД MS SQL 2008.

Вихідні дані представляють собою результати роботи алгоритму, а саме результати виконання кроків алгоритму та підсумки його роботи.

У структурі рекомендаційної системи можна виділити наступні модулі:

- Модуль входу в систему;
- Модуль вибору моделі прогнозування;
- Модуль вибору методу прогнозування
- Модуль вибору метрики подібності;
- Модуль прогнозування рекомендацій;
- Модуль оцінки точності прогнозування;
- Модуль візуалізації результатів.

Модуль прогнозування рекомендацій та модуль оцінки точності прогнозування мають двохсторонній зв'язок з БД профілів користувачів і предметів.

Структурна схема рекомендаційної системи зображена на рис. 4.1.

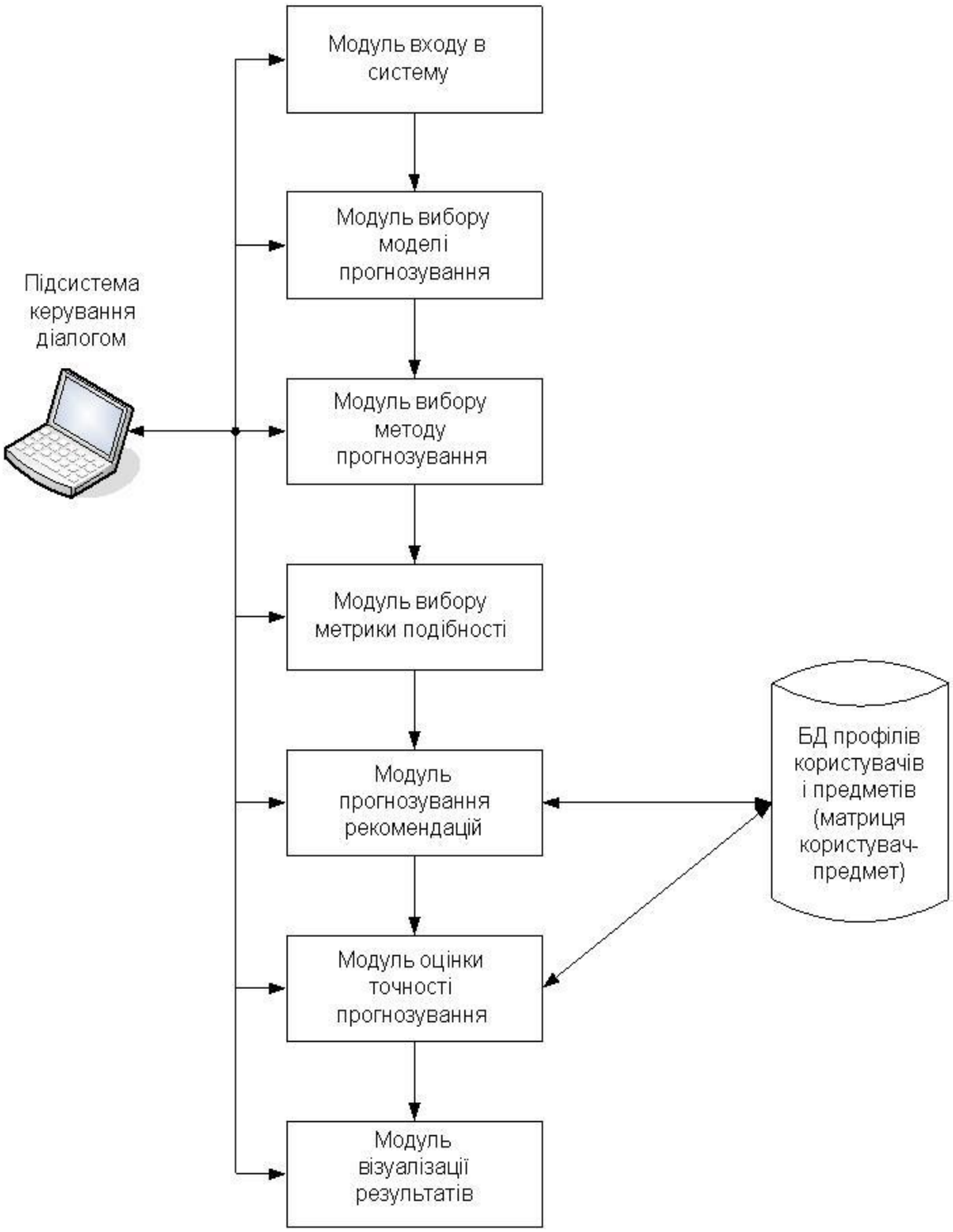


Рис.4.1 – Структурна схема рекомендаційної системи

4.3. Структура програмного забезпечення рекомендаційної системи

Структура програмного забезпечення рекомендаційної системи відображає модулі програмного забезпечення прогнозування рекомендацій та зв’язки між ними та зображена на рис. 4.2.

У програмному забезпеченні можна виділити наступні модулі:

- Модуль прогнозування рекомендацій за допомогою асоціативних правил;
- Модуль прогнозування рекомендацій для груп користувачів;
- Модуль прогнозування рекомендацій методом колаборативної фільтрації;
- Модулі моделей прогнозування рекомендацій для груп користувачів;
- Модулі розрахунку коефіцієнтів подібності для колаборативної фільтрації;
- Модулі розрахунку точності прогнозування рекомендацій.

4.4. Структура класів програмного забезпечення рекомендаційної системи

Головним класом програми є клас Form1, в ньому реалізовано основні функції, які забезпечують користувацьких інтерфейс. Цей клас на високому рівні абстракції використовує екземпляри інших класів: класи користувачів, фільмів (User і Film) і алгоритмів надання рекомендацій (Recomendation). Також в ньому реалізовано основні функції введення та виведення даних, їх візуалізацію у табличному вигляді.

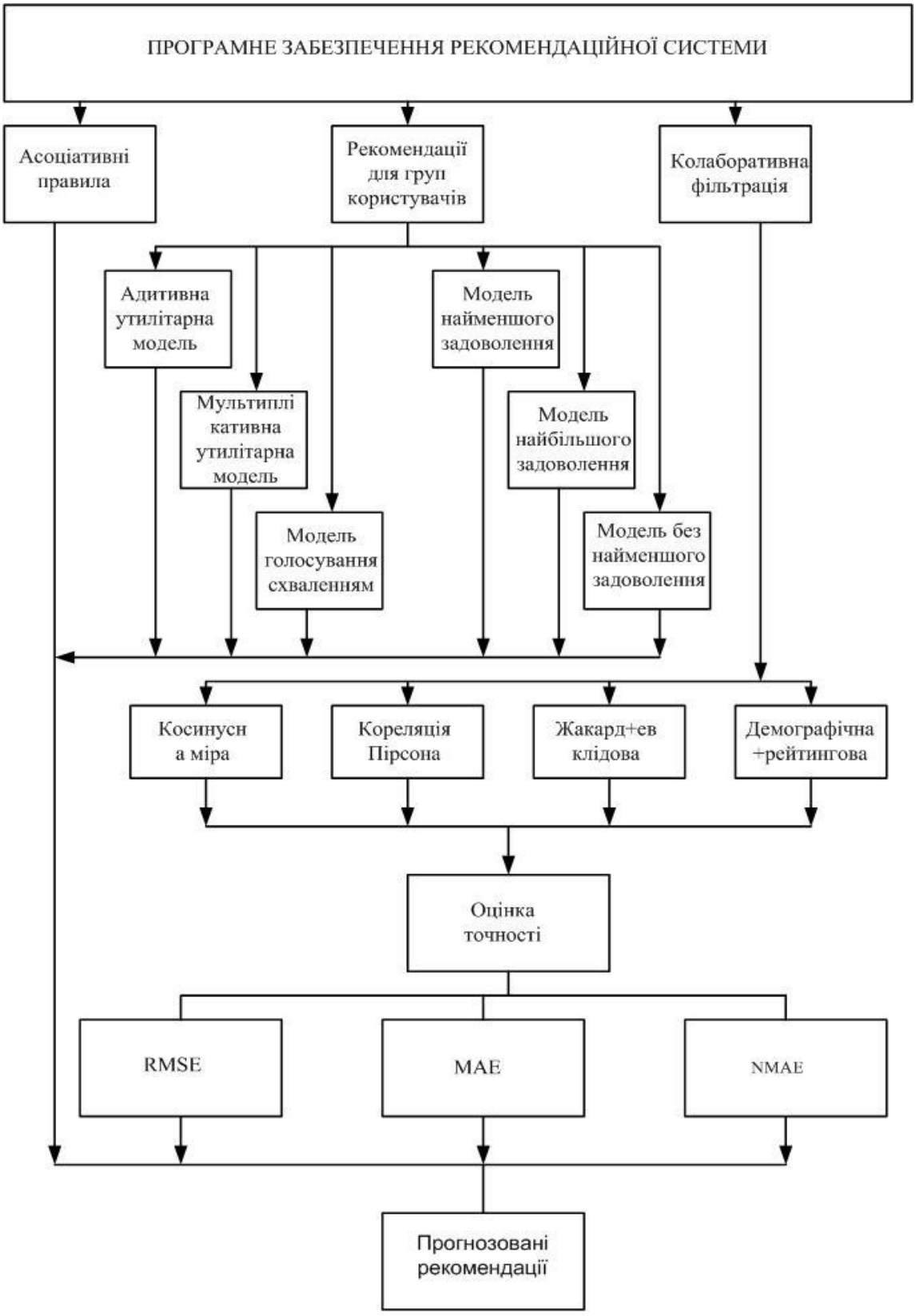


Рис.4.2 – Структура програмного забезпечення прогнозування рекомендацій рекомендаційної системи

Клас Film містить методи:

1. matchFilms – знаходить коефіцієнт схожості фільмів за коефіцієнтом Пірсона. Приймає вхідний параметр типу Film, повертає значення в діапазоні від 0 до 1, чим значення вище, тим більш схожі фільми. Приклад виклику this.matchFilms(Filmf).

Поля:

1. id – цілочисельне значення, ідентифікатор фільму;
2. title – текстове значення назви
3. rating – дійсне значення, оцінка фільму;
4. userId – цілочисельне значення, ідентифікатор користувача що поставив оцінку;
5. marks – масив дійсних значень рейтингів кожного жанру;.
6. genre – масив жанрів фільму;
7. match – схожість фільму.

Клас User містить методи:

1. addFilm – додає фільм до списку переглянутих користувачем фільмів. Приймає вхідний параметр типу Film. Не повертає ніякого значення, але додає елементи до списку films даного об’єкта. Приклад виклику this.addFilm(Filmf);
2. fillFilm – додає до списку переглянутих фільмів всі фільми з БД які оцінив користувач. Має 2 перезавантажених варіанти виклику: 1) 1 вхідний параметр типу movielensEntities (об’єкт класу ObjectContext), не повертає значення; 2) 2 вхідних параметра: 1- ObjectContext; 2- типу double-коефіцієнт схожості до заданого користувача. Приклади виклику:

this.fillFilms(movielensEntities c), this.fillFilms(movielensEntities c, double match);

3. manhatten – знаходить коефіцієнт схожості за Манхетенівською відстанню. Приймає 4 вхідні параметри: 1- типу long – ідентифікатор користувача для якого потрібно порахувати Манхетенівську відстань, 2 –movielensEntities – підключення до БД, 3- Dictionary<long,long[]>[] – масив словників, ключами яких є ідентифікатори користувачів, а значеннями – ідентифікатори фільмів, що є спільні для поточного користувача і користувача з заданим ідентифікатором, 4- типу int – індекс в масиві словників, що відповідає за словник для Манхетенівської відстані. Повертає значенни типу double. Приклад виклику: double match = currentUser.manhatten(user.user_id, context, mainForm.CommonFilms, i);
4. Evklide - знаходить коефіцієнт схожості за Евклідовою відстанню. Приймає 4 вхідні параметри: 1- типу long – ідентифікатор користувача для якого потрібно обчислити Манхетенівську відстань, 2 –movielensEntities – підключення до БД, 3- Dictionary<long,long[]>[] – масив словників, ключами яких є ідентифікатори користувачів, а значеннями – ідентифікатори фільмів, що є спільні для поточного користувача і користувача з заданим ідентифікатором, 4- типу int – індекс в масиві словників, що відповідає за словник для Манхетенівської відстані. Повертає значенни типу double. Приклад виклику: doublematch = currentUser.evklide(user.user_id, context, mainForm.CommonFilms, i);
5. pearson - знаходить коефіцієнт схожості за коефіцієнтом схожості Пірсона. Приймає 4 вхідні параметри: 1- типу long – ідентифікатор користувача для якого потрібно порахувати Манхетенівську відстань, 2 –movielensEntities – підключення до БД, 3- Dictionary<long,long[]>[] – масив словників, ключами яких є

ідентифікатори користувачів, а значеннями – ідентифікатори фільмів, що є спільні для поточного користувача і користувача з заданим ідентифікатором, 4- типу int – індекс в масиві словників, що відповідає за словник для Манхетенівської відстані. Повертає значенни типу double. Приклад виклику: double match = currentUser.pearson(user.user_id, context, mainForm.CommonFilms, i);

6. count – знаходить кількість спільно оцінених фільмів між користувачами. Приймає 2 вхідних параметри: 1- типу long – ідентифікатор користувача, кількість спільних з яким потрібно знайти і movielensEntities – підключення до БД. Повертає значення типу int. Приклад виклику intcount = currentUser.count(user.user_id, mainForm.Context).

Поля:

1. id – цілочисельне значення, ідентифікатор користувача;
2. match – дійсне значення, коефіцієнт схожості;
3. films – спикок об’єктів типу Film, список фільмів, оцінених даним користувачем.

Клас Recomendation містить методи:

1. simpleRecomendation – знаходить прогнозовані оцінки фільмів з рекомендацією по користувачах. Приймає 3 вхідні параметри: 1 - типу DataGridView – таблиця з ідентифікаторами користувачів, коефіцієнтами схожості на заданого та кількістю спільних фільмів, 2 - типу movielensEntities – підключення до БД, 3 – типу User – поточний користувач системи;
2. weightedSumRecomendation - знаходить прогнозовані оцінки фільмів з рекомендацією по користувачах. Приймає 3 вхідні параметри: 1 - типу DataGridView – таблиця з ідентифікаторами користувачів, коефіцієнтами схожості на заданого та кількістю

спільних фільмів, 2 - типу movielensEntities – підключення до БД, 3 – типу User – поточний користувач системи;

3. findWeightCoef – знаходить коефіцієнт схожості користувачів по відповідному методу (косинусна відстань, манхетенівська відстань, евклідова відстань і коефіцієнт кореляції Пірсона).

Поля в даному класі відсутні.

Клас Form1 містить методи:

1. Form1_CheckedChanged – обробник вибору методу обрахунку коефіцієнта схожості користувача.
2. simpleRecomendationMenuItem_Click – обробник вибору пункту меню Проста рекомендація. Не повертає ніякого значення.
3. Вихід ToolStripMenuItem_Click – обробник вибору пункту меню Вихід.
4. weigedSumРекомендаціяToolStripMenuItem_Click – обробник вибору пункту меню Зважена сума.
5. СпрощенаРегресіяToolStripMenuItem_Click – обробник вибору пункту меню Спрощена регресія.
6. Form1_Load – обробник події Load форми.
7. CurrentUser – властивість, яка повертає поточного користувача.
8. DestroyForm – метод, призначений для коректного закриття дочірніх форм.
9. UsersMatch – масив коефіцієнтів схожості користувачів за відповідними методами.
10. CommonFilms – метод, який повертає масив словників ключами в яких є ідентифікатори користувачів, а значеннями – масиви ідентифікаторів спільних фільмів.
11. CheckedCheckboxes – метод, який повертає вибрані методи обрахунку коефіцієнту схожості користувачів.

12. Context – властивість, яка повертає об’єкт, за допомогою якого здійснюється підключення до БД.

Поля:

1. checkedCheckboxes – цілочисельне поле, де зберігаються вибрані методи образунку коефіцієнту схожості користувачів шляхом встановлення відповідного біту в 0 чи 1.
2. forms – масив типу Form, який зберігає відкриті дочірні форми програми.
3. usersMatch – двовимірний масив типу UserMatch, призначений для збереження коефіцієнтів схожості користувачів.
4. currentUser – поле типу User, призначене для збереження поточного користувача.
5. commonFilms – масив типу Dictionary <long, long[]>, призначений для збереження спільних фільмів.
6. context – поля для збереження об’єкту, за допомогою якого здійснюється підключення до БД.

Клас RecomendedFilm містить автоматичні властивості:

1. FilmID – ідентифікатор фільму, який рекомендується користувачу для перегляду.
2. ForecastMark – прогнозована оцінка.
3. FilmTitle – назва фільму, який рекомендується користувачу для перегляду.

Поля в даному класі генеруються автоматично.

Клас UserMatch реалізує інтерфейс IComparable <UserMatch> і містить наступні методи:

1. UserID – ідентифікатор користувача для якого розраховані коефіцієнти подібності.
2. Match – коефіцієнт подібності.
3. FilmsCount – кількість спільних фільмів.
4. CompareTo(UserMatch other) – метод для порівняння двох об’єктів типу UserMatch.

Поля в даному класі генеруються автоматично.

Діаграма основних класів представлена на рис.4.3

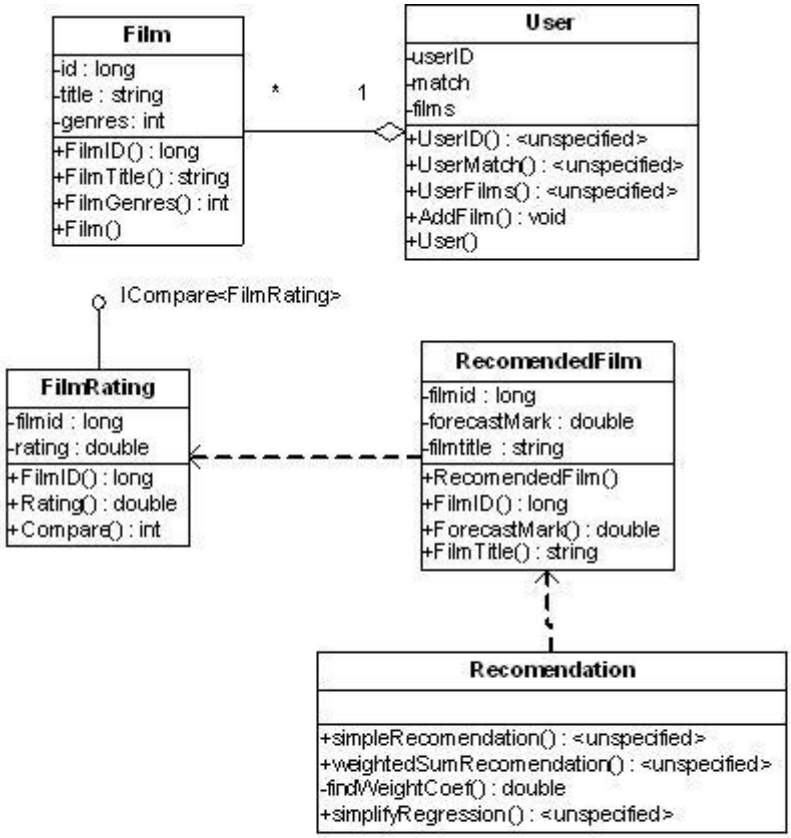


Рис.4.3 – Діаграма основних класів

Схема даних розробленої БД зображена на рис. 4.4

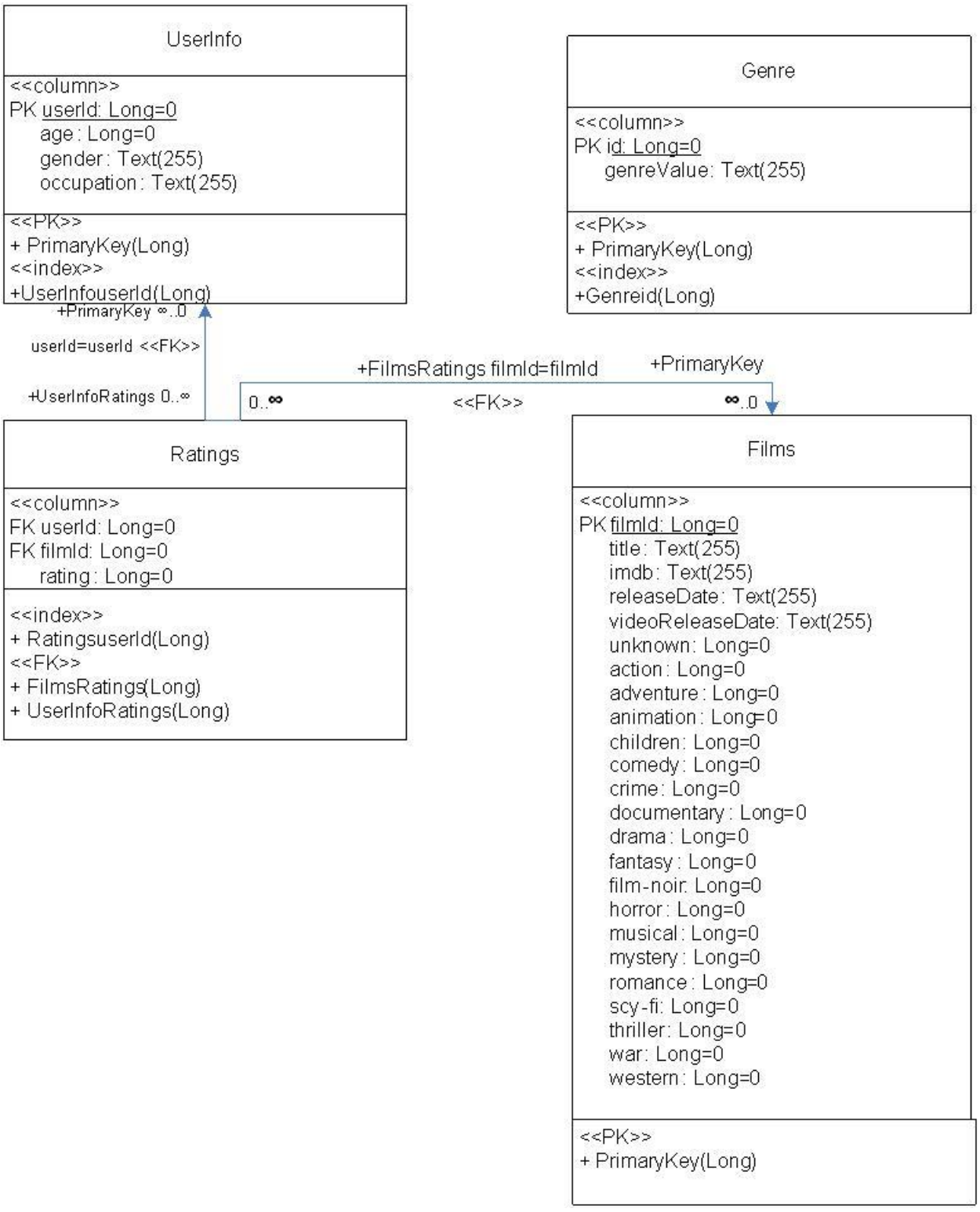


Рис. 4.4 – Схема даних розробленої БД

4.5. Інформаційне забезпечення рекомендаційної системи

У структуру тестової реляційної БД MovieLens [113] входять наступні таблиці: **users, age_ranges, genders, occupations, zipcodes, ratings, movies,**

movie_genres, genres. Модель БД в концептуальній формі представлена на рис.4.5.

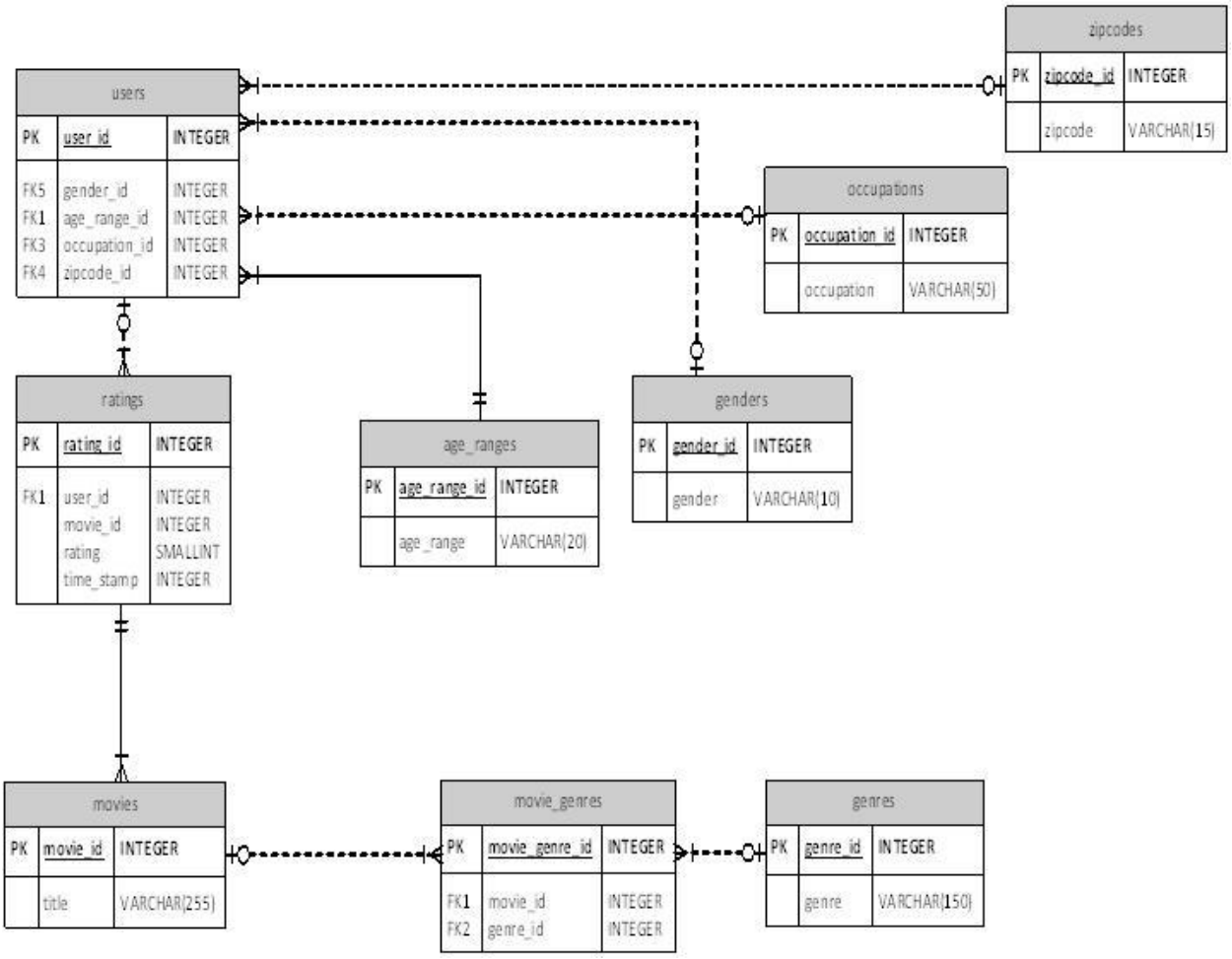


Рис.4.5 – Концептуальна модель БД

Таблиця **users** містить у собі інформацію про ідентифікатор користувача (user_id), інформацію про ідентифікатор статі (gender_id), інформацію про ідентифікатор вікового діапазону (age_range_id), інформацію про ідентифікатор роду заняття (occupation_id) та інформацію про ідентифікатор поштового індексу(zipcode_id).

Таблиця **age_ranges** містить інформацію про ідентифікатор вікового діапазону (age_range_id) та інформацію про самий віковий діапазон (age_range).

Таблиця **genders** складається з полів gender_id (ідентифікатор статі) та gender (назва статі).

Таблиця **occupations** містить у собі поля occupation_id, яке містить інформацію про ідентифікатор роду заняття та сама назва роду заняття – occupation.

Таблиця **zipcodes** містить zipcode_id (ідентифікатор поштового індексу) та zipcode (поштовий індекс).

Таблиця **ratings** містить у собі інформацію про рейтинги, які поставив користувач фільму, та складається з полів rating_id (ідентифікатор рейтингу), user_id (ідентифікатор користувача), movie_id (ідентифікатор фільму), rating (оцінка фільму), time_stamp (мітка фільму).

Таблиця **movies**містить у собі інформацію про ідентифікатор фільму (movie_id) та інформацію про назву фільму (title).

Таблиця **movie_genres** складається з полів movie_genre_id (ідентифікатор жанру фільму), movie_id (ідентифікатор фільму) та genre_id (ідентифікатор жанру).

Таблиця **genres** містить у собі інформацію про жанри та складається з полів genre_id (ідентифікатор жанру) та genre (назва жанру)

БД MovieLensмістить у собі інформацію про **6040 користувачів, 3900 фільмів, 1000209 поставлених користувачами оцінок фільмам.**

Виходячи з цих даних, можна порахувати розрідженість БД. Розрідженість матриці користувач-предмет можна обчислити за формулою:

$$SP = 1 - \frac{nR}{nUSER * nITEM} = 1 - \frac{1000209}{6040 * 3900} = 0{,}9576\ . \qquad (4.1)$$

4.6. Розділення тестової матриці користувач-предмет на розрахункову та тестові множини

4.6.1. Можливість вибору методу розділення

Існують три методи розділення можливості:

- **Просте розділення** (Simple split) – розділяє дані на дві множини, навчальну та тестову. Користувач може сам вибрати пропорційний розмір тестової частини.
- **Заздалегідь визначене розділення** (Predefined split) – використовується для перевірки правильності роботи реалізованого алгоритму.
- **К-кратна крос-валідація** (K-fold Crossvalidation) – набір даних ділиться на К частини однакового розміру, після цього проводяться експерименти К разів з кожною К-тою частиною в ролі тестових даних і рештою як навчальними даними. Програма показує результати оцінювання для кожної стадії окремо.

4.6.2. Принцип роботи з заздалегідь визначеним розділенням

Існує можливість використання файлу з заздалегідь визначеним розділенням для спрощення повторних експериментів. Завдяки цій опції, використовуючи однакові дані, розробники та тестувальники алгоритмів можуть оцінити свої алгоритми. Для початку необхідно мати правильно підготовлений файл з інформацією про розділення. Далі при запуску програми вибирається опція використання заздалегідь визначеного розділення та вибирається назва цього файлу. Існує два способи створення цього файлу: вручну та автоматично.

- **Розділення вручну**: формат файлу простий. В кожному рядку зберігається кожна тестова точка у вигляді: ідентифікатор

користувача – ідентифікатор предмету. Елементи розділяються пробілом.

- **Автоматичне розділення**: завдяки цьому можна розділити заданий набір даних з заданим співвідношенням навчальної та тестової частин.

4.7. Оцінка точності

Якість РС оцінюється за результатами тестування. Тип метрик, який використовується, залежить від типу КФ додатків. Згідно [41,57], метрики оцінки РС можна поділити на такі основні категорії: прогнозуюча метрика точності, такі як САП і його варіації; метрики точності класифікації, такі, як точність, відклик, F1-міра, і чутливості ОПХ, ранг показників точності.

4.7.1. Середня абсолютна похибка і нормавана середня абсолютна похибка

Найбільш широко використовуються метрики в КФ науковій літературі САП, яка обчислює середню величину абсолютної різниці між передбаченнями і істинним оцінками.

$$MAE = \frac{\sum_{i,j} \left| p_{ij} - \bar{r}_{ij} \right|}{n}, \tag{4.2}$$

де n є загальна кількість рейтингів усіма для всіх користувачів, p_{ij} – прогнозом рейтингу для користувача і на предмет j, і $\bar{r}_{ij}$ – середнє значення рейтингів i-того користувача для всіх предметів. Чим нижче САП, тим кращий прогноз.

Різні РС можуть використовувати різні чисельні шкали. НСАП, нормалізує САП, щоб виразити похибки як відсотки повного масштабу:

$$NMAE = \frac{MAE}{r_{\max} - r_{\min}}, \tag{4.3}$$

де r_{max} і r_{min} – верхні і нижні межі оцінок.

4.7.2. Коренева середньо квадратична похибка

Іншою популярною метрикою контролю точності є КСКП:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i,j}(p_{ij} - r_{ij})^2} \ , \qquad (4.4)$$

де n − загальна кількість рейтингів усіх користувачів, p_{ij} − прогноз рейтингу користувача i на предмет j, і r_{ij} − фактичне значення рейтингу. КСКП посилює значення абсолютних похибок між прогнозованими і дійсними значеннями.

4.8. Блок-схеми алгоритмів для методів рекомендаційної системи

На рис.4.6 наведено блок-схему алгоритму прогнозування рекомендацій методом колаборативної фільтрації.

На рис. 4.7 – 4.10 наведені блок-схеми алгоритмів розрахунку мір подібності для векторів профілів користувачів і предметів.

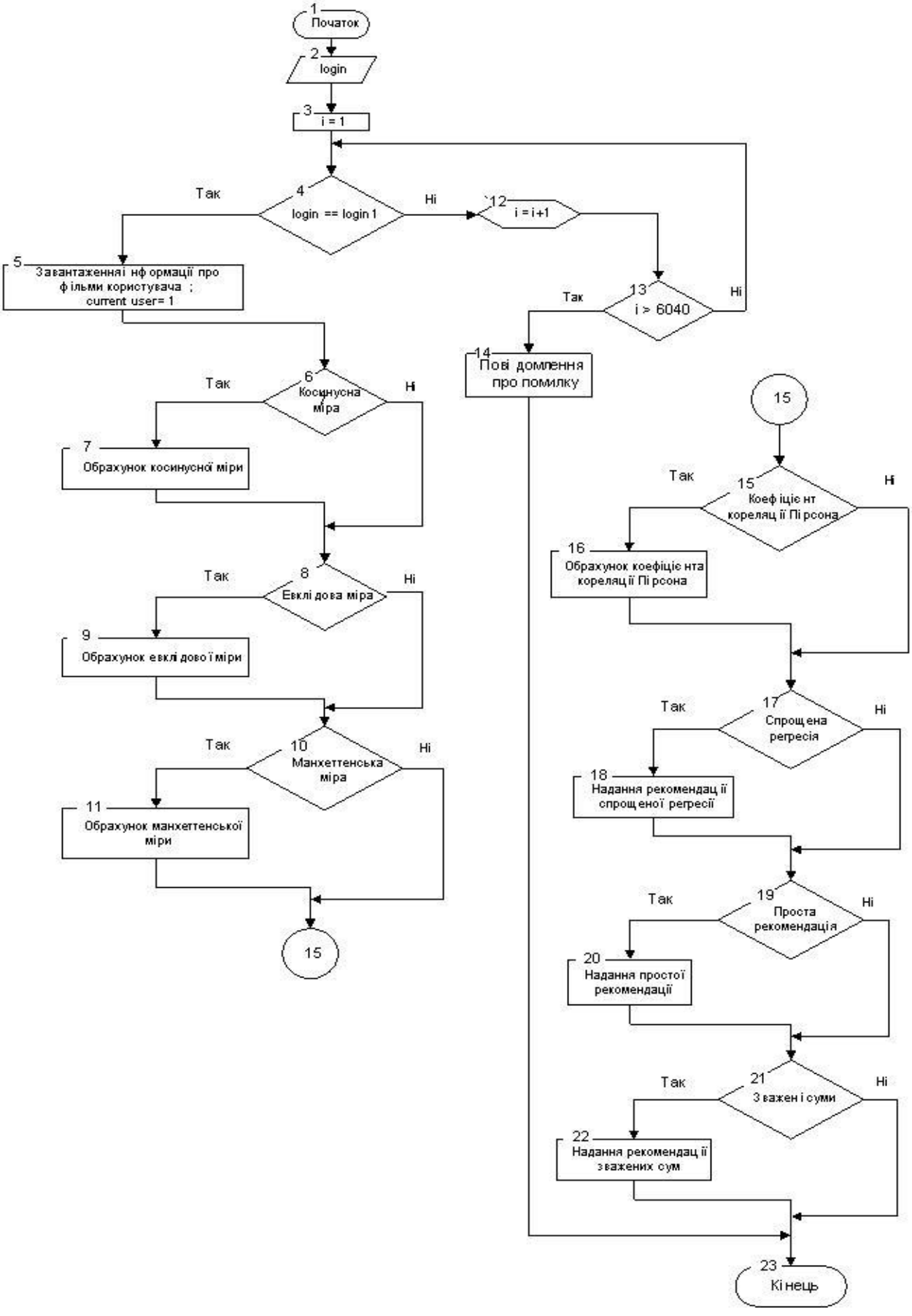


Рис.4.6 – Блок-схема алгоритму прогнозування рекомендацій методом колаборативної фільтрації

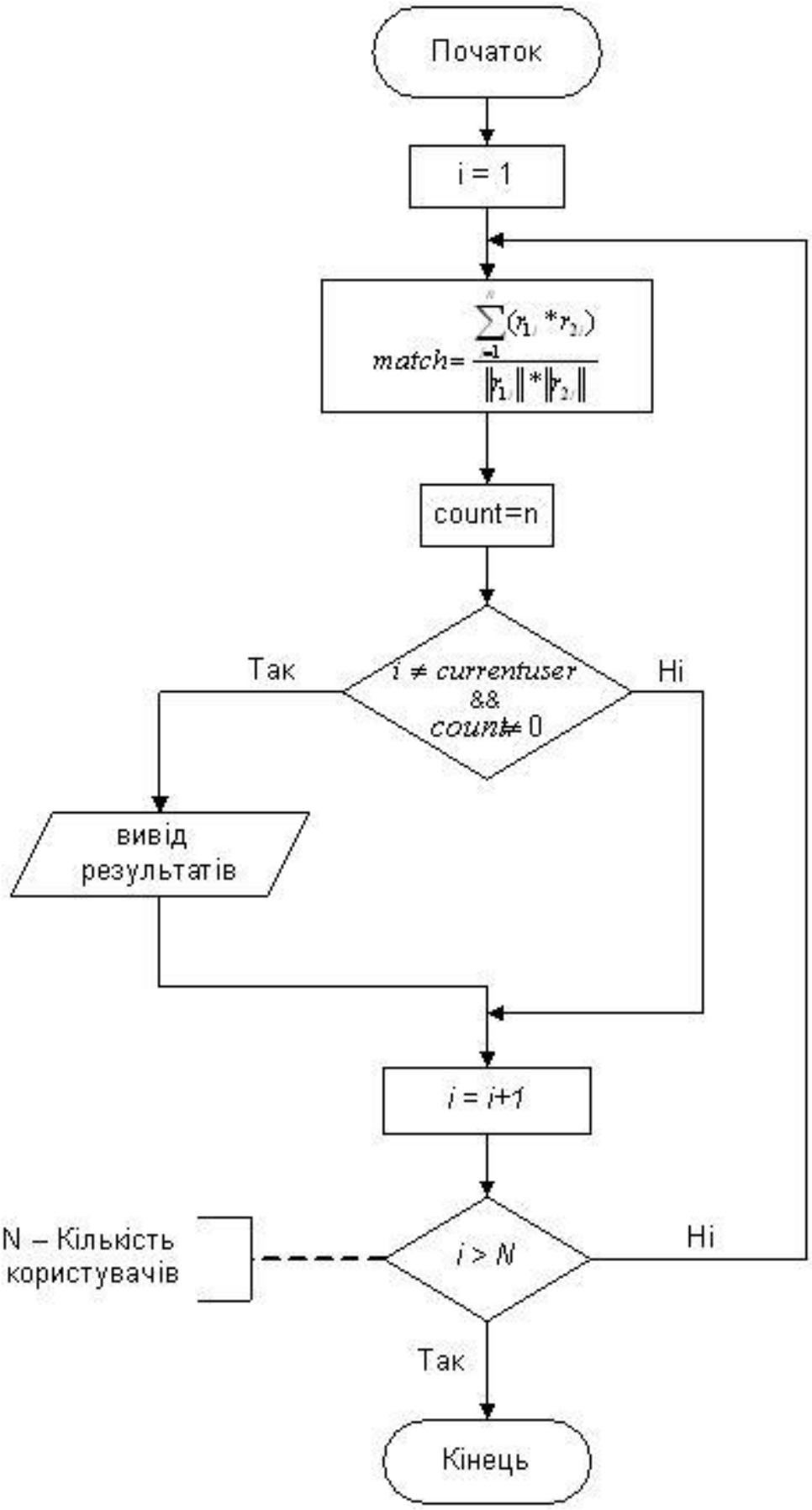


Рис.4.7 – Блок-схема алгоритму знаходження косинусної міри подібності

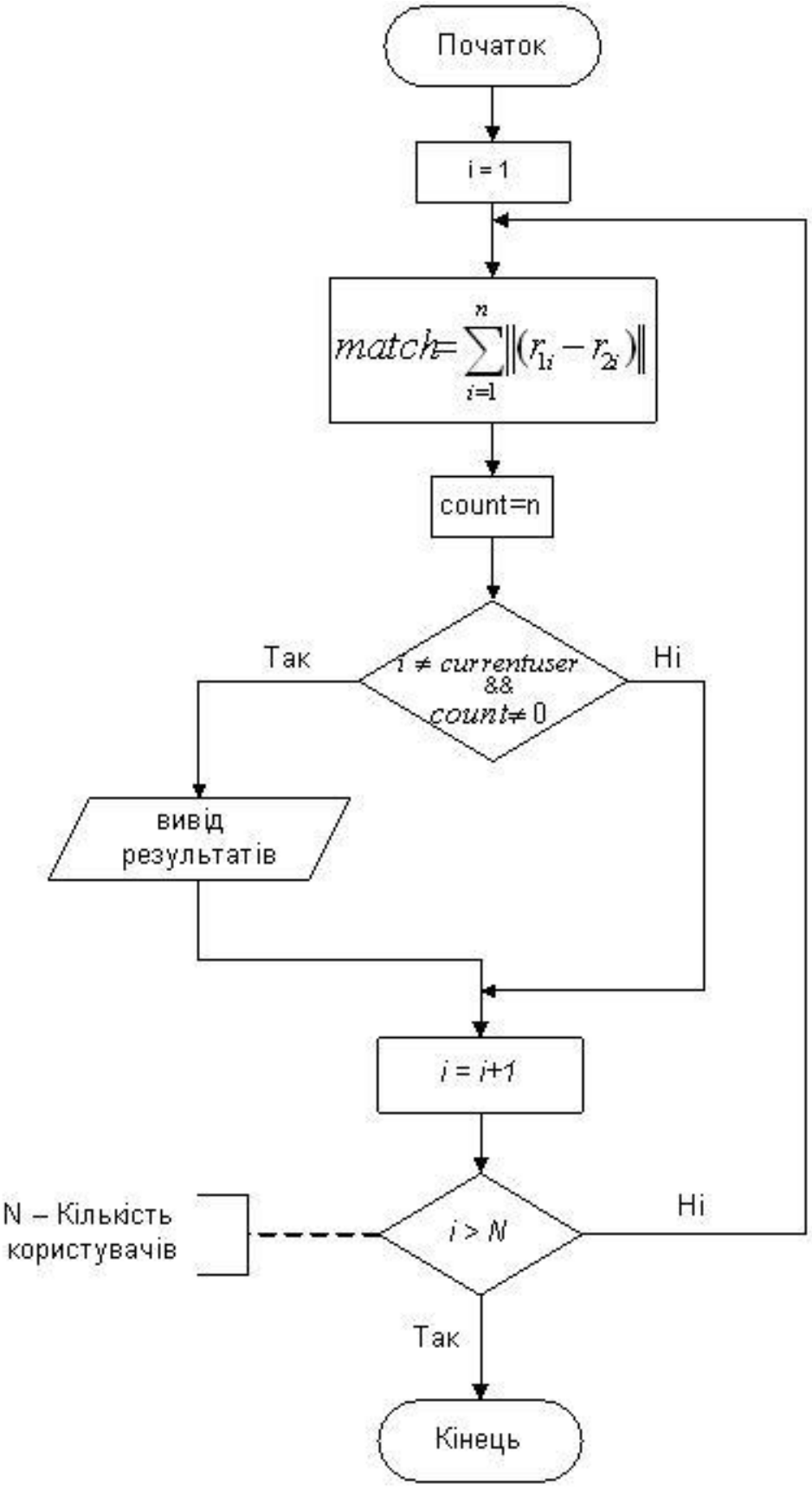


Рис.4.8 – Блок-схема алгоритму знаходження Манхеттенської міри подібності

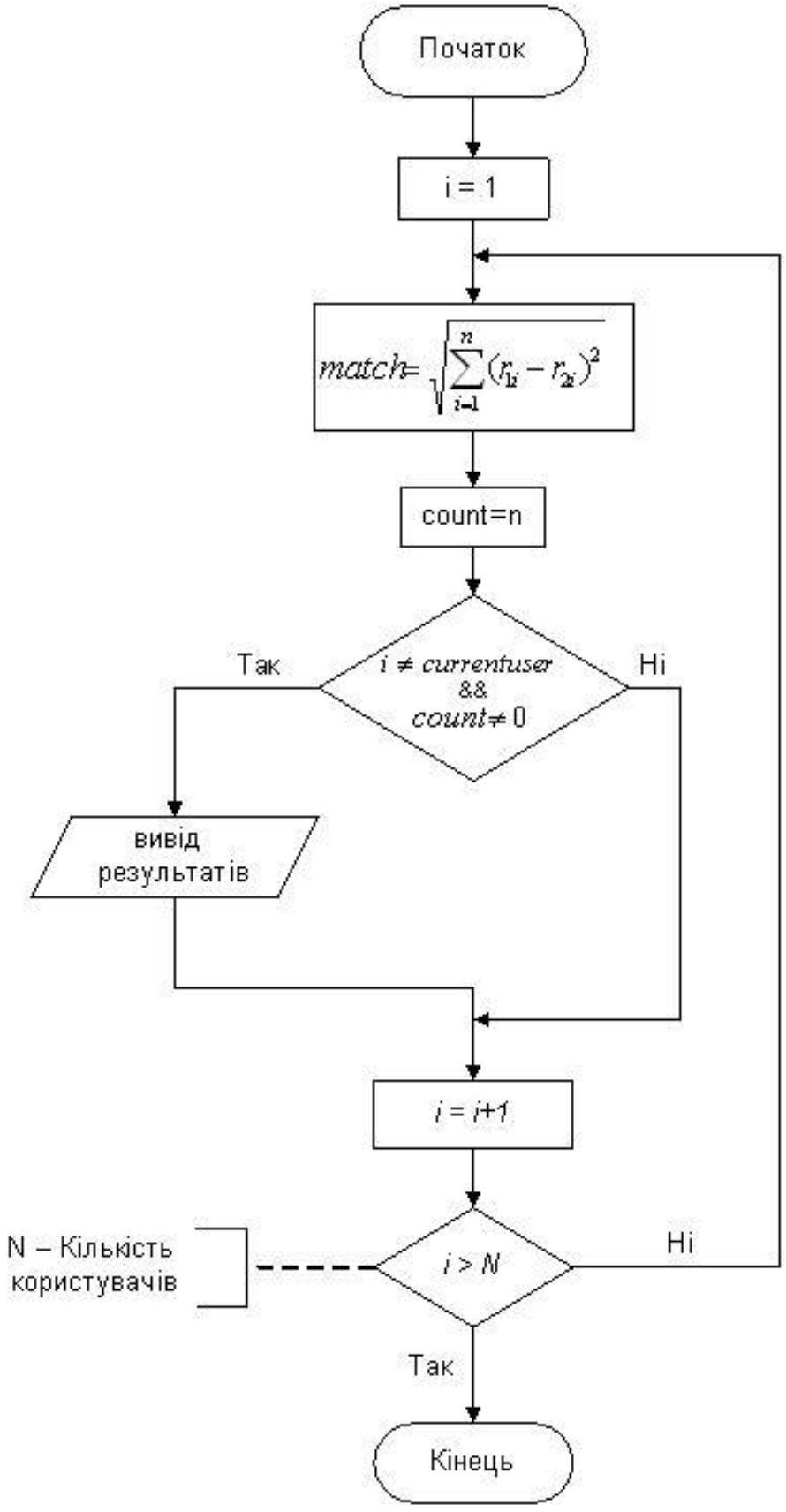


Рис.4.9 – Блок-схема алгоритму знаходження евклідової міри подібності

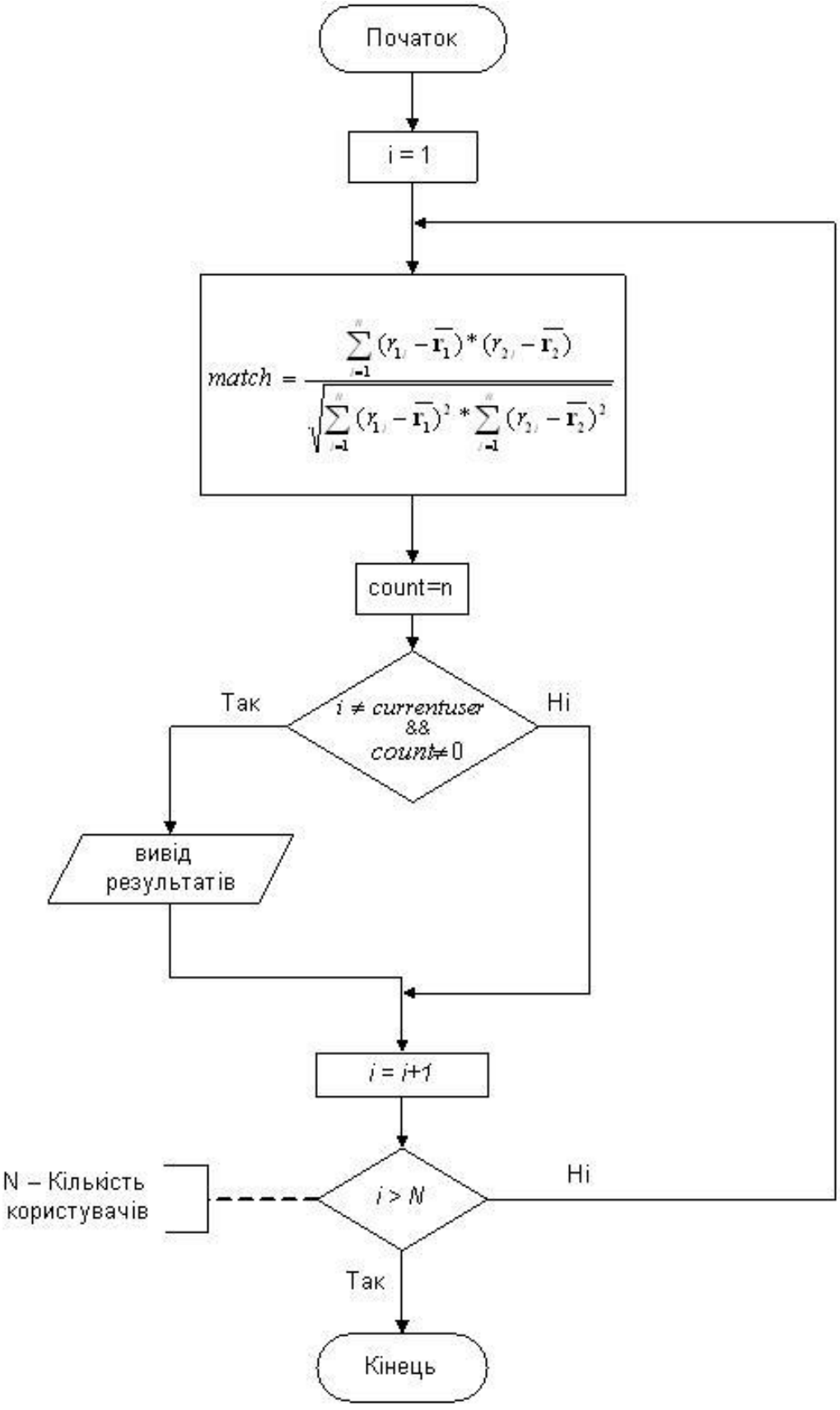


Рис.4.10 – Блок-схема знаходження коефіцієнта кореляції Пірсона

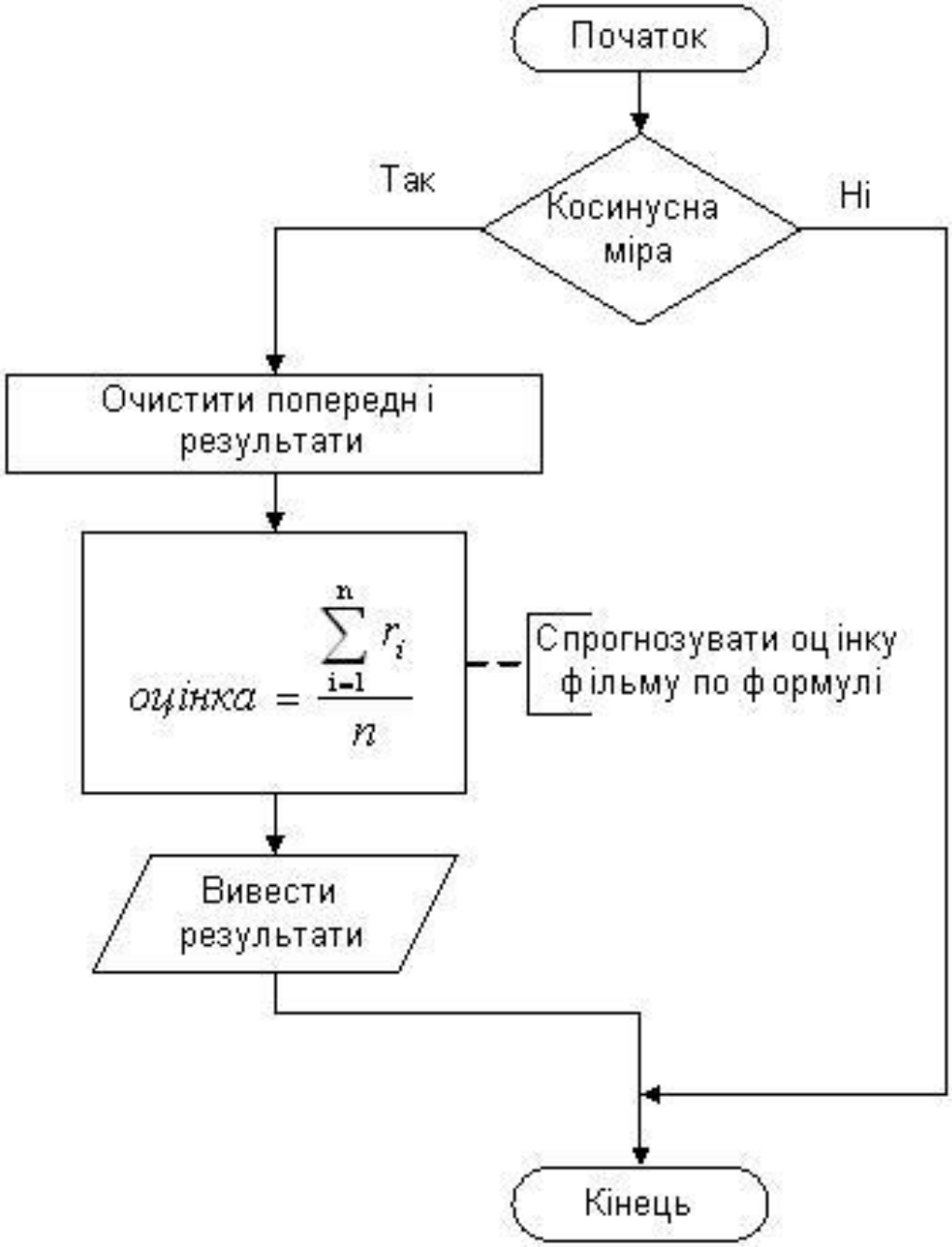


Рис.4.11 – Блок-схема надання простої рекомендації

На рис. 4.11 наведена блок-схема алгоритму простої рекомендації.

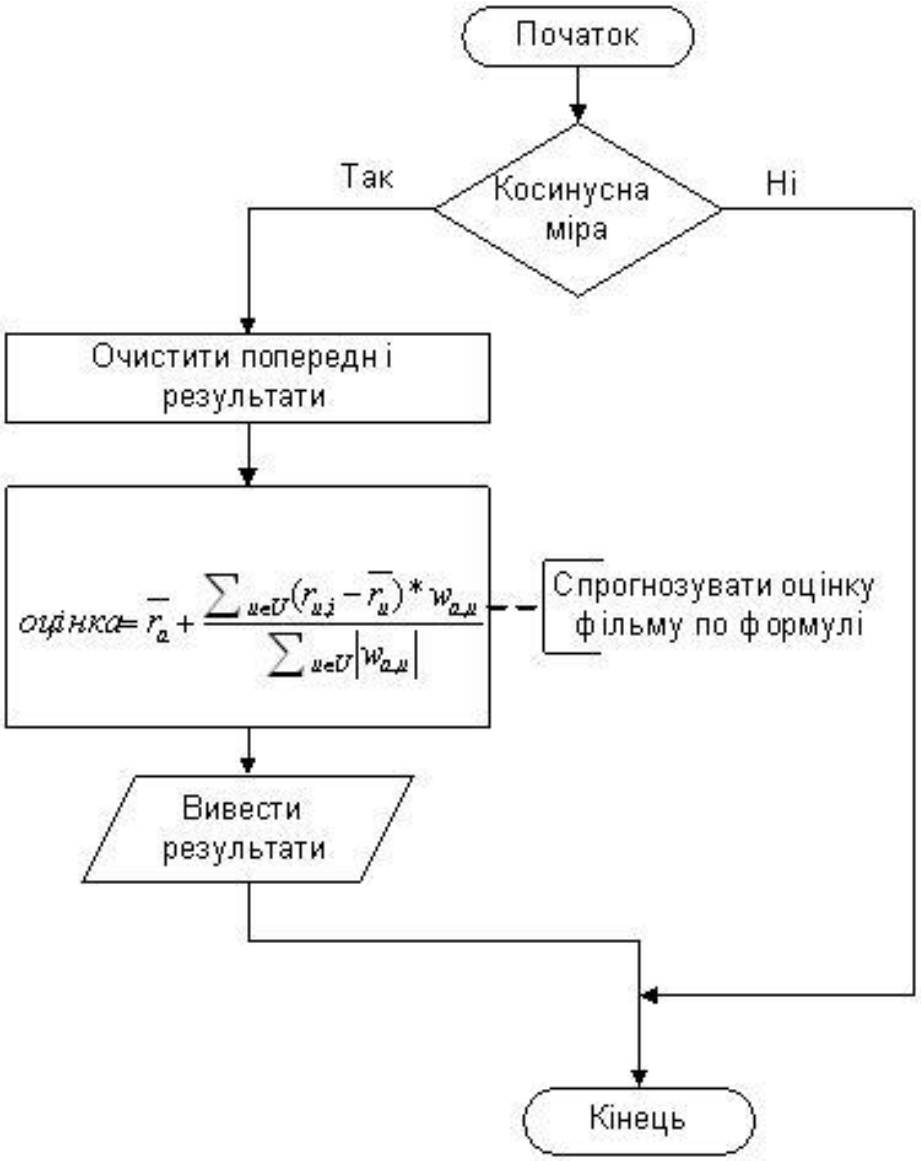


Рис.4.12 – Блок-схема надання рекомендації методом зваженої суми

На рис. 4.12 представлена частина блок схеми алгоритму прогнозування рекомендації методом зваженої суми. Вагові коефіцієнти обчислюються в залежності від вибраного методу розрахунку міри подібності.

На рис. 4.13 представлена блок схема оцінки точності по MAE. Блок-схема алгоритму, приведеного на рис.4.13 справедлива для прогнозування рекомендації методом простої рекомендації, лінійної і спрощеної лінійної регресії, для методу колаборативної фільтрації і гібридних методів прогнозування рекомендацій.

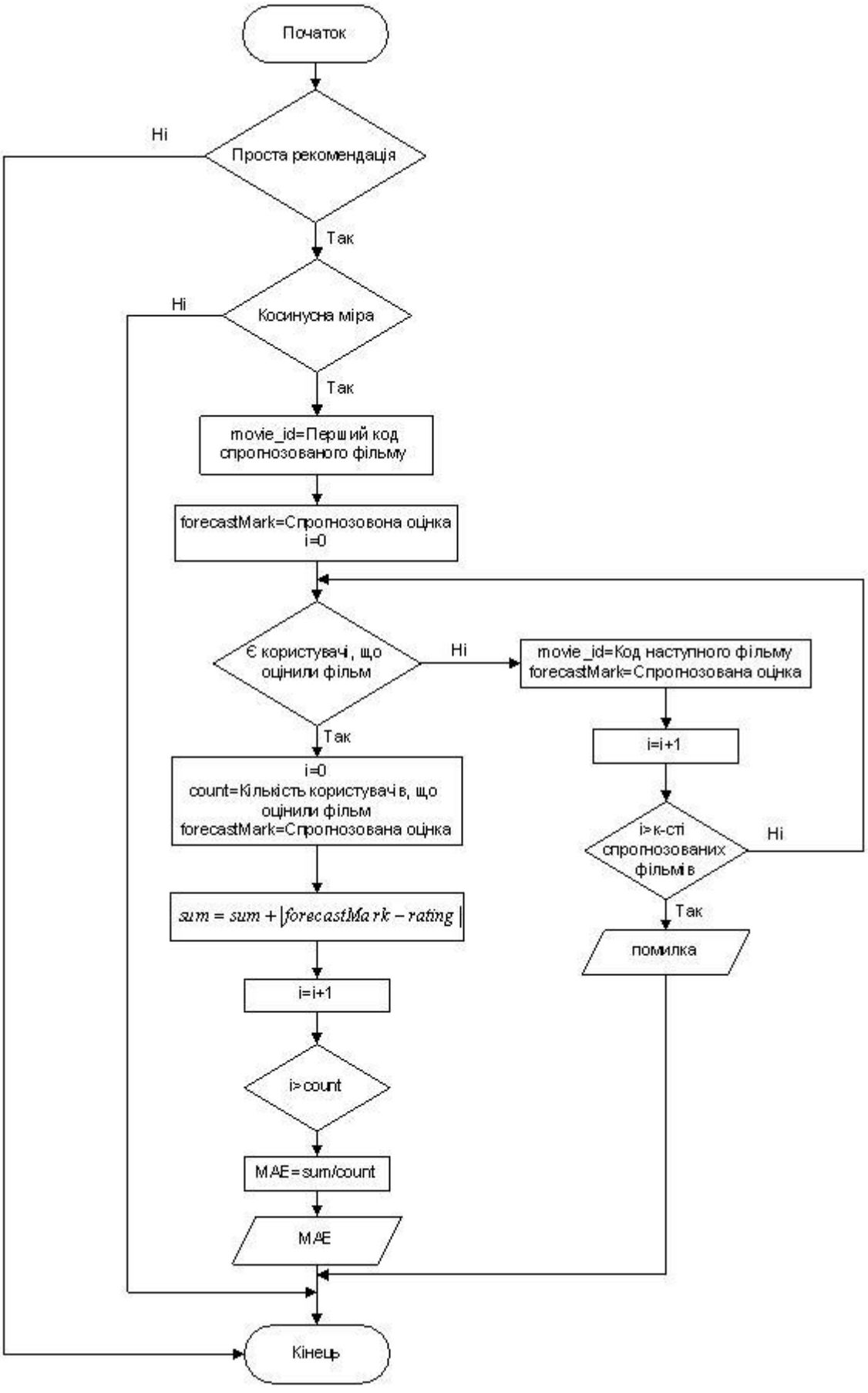


Рис.4.13 – Блок-схема оцінки точності по MAE

4.9. Результати тестування на наборі даних MovieLens

Як показали результати тестування, Евклідова міра подібності дає найгірший результат і не є ефективною для методу зваженої суми. Найбільш ефективною мірою є запропонована в розділі 2 комбінована міра (Евклідова+Жаккард) (рис. 4.14).

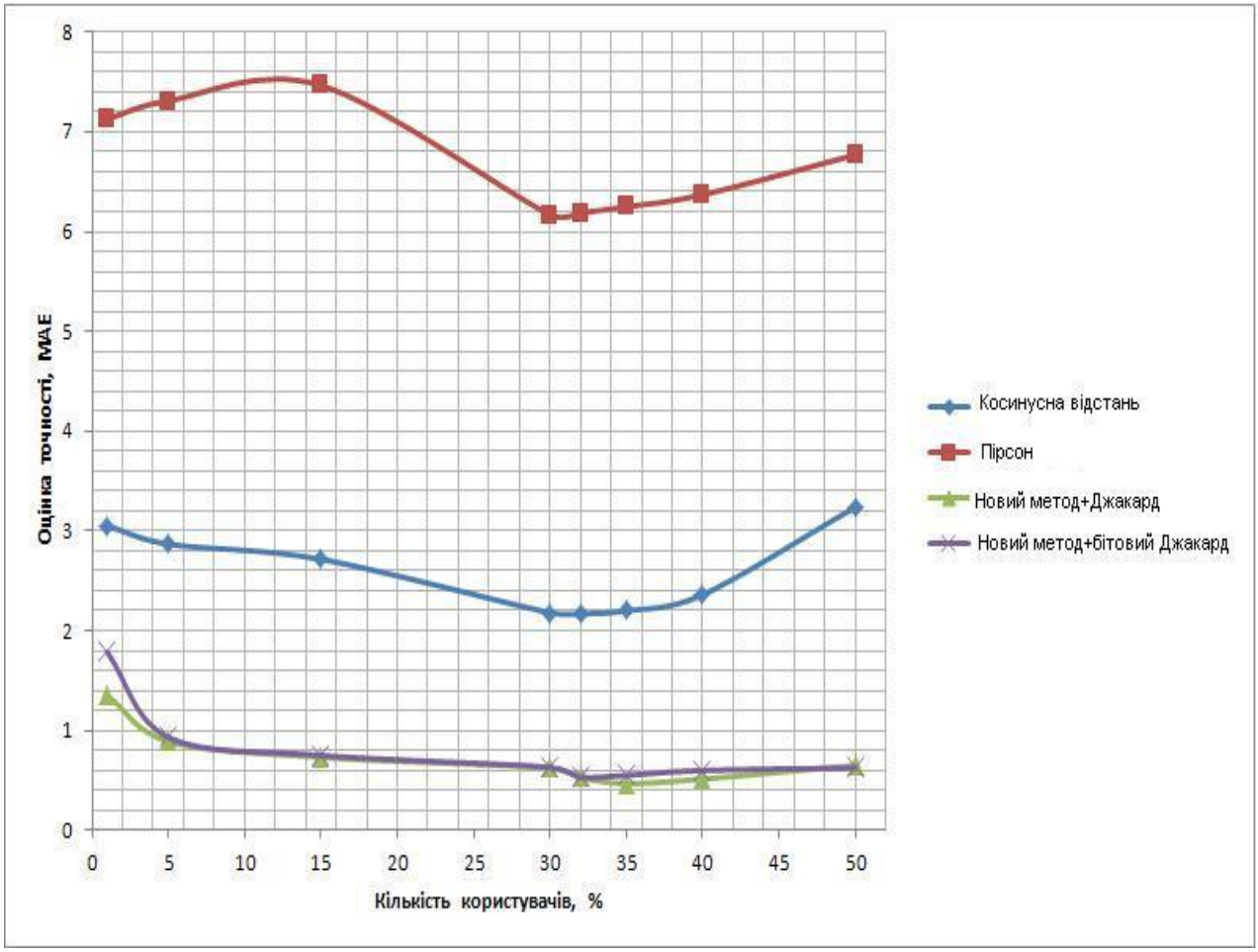


Рис. 4.14 – Результати оцінки точності прогнозування методом зваженої суми по MAE

При тестуванні методу простої рекомендації результати тестування показали, що усі міри подібності дали одинаковий результат (рис. 4.15). Це є вірним, оскільки проста рекомендація не використовує коефіцієнти мір подібності. Процес надання цієї рекомендації є найпростішим і найшвидшим.

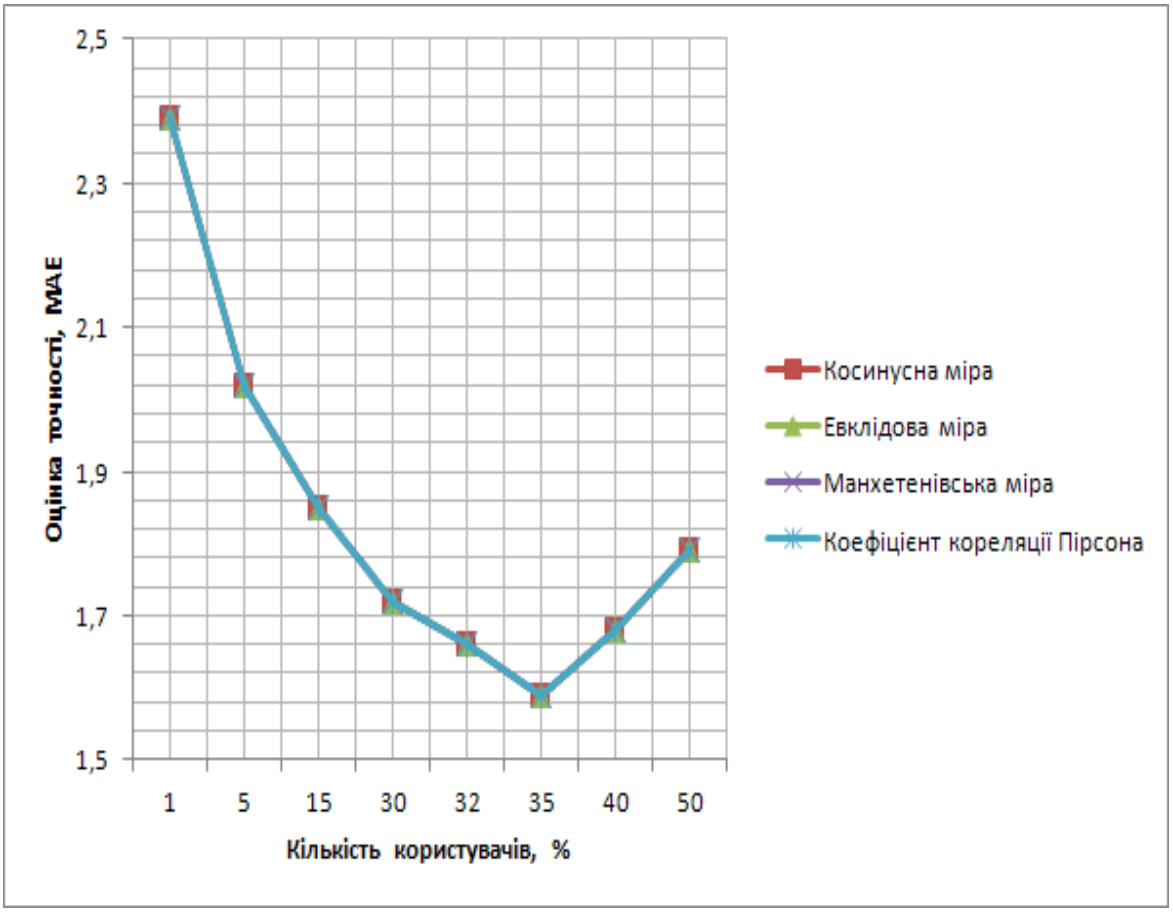


Рис. 4.15 – Результати оцінки точності прогнозів методом простої рекомендації по MAE

На рис. 4.16 наведено результати тестування точності прогнозування рекомендацій в залежності від кількості подібних векторів профілів в околі активного користувача для методу зваженої суми для тестового набору MoviLens при використанні косинусного коефіцієнта подібності і демографічно-рейтингового коефіцієнта подібності. По осі координат відкладено абсолютну похибку прогнозування (MAE).

UB1− міра подібності на основі оберненої евклідової відстані;

UB2 − модифікована міра подібності, домінуюче значення демографічної подібності;

UB3 − модифікована міра подібності, домінуюче значення рейтингової подібності.

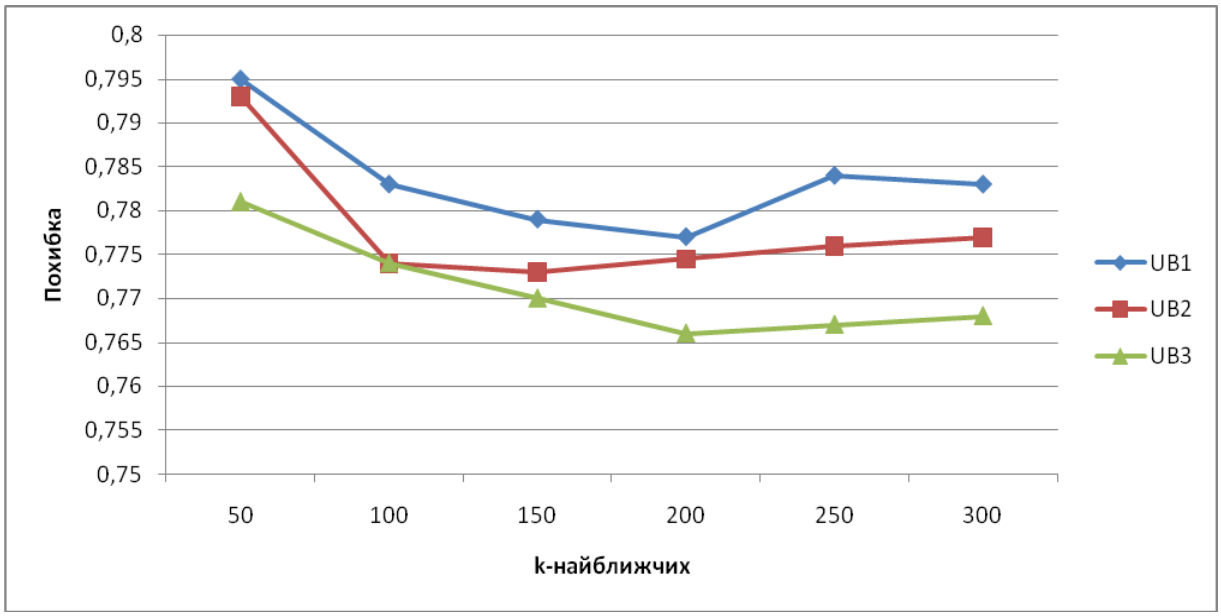


Рис. 4.16 – Залежність похибки прогнозування рекомендацій для тестового набору MoviLens в залежності від розміру області k-найближчих

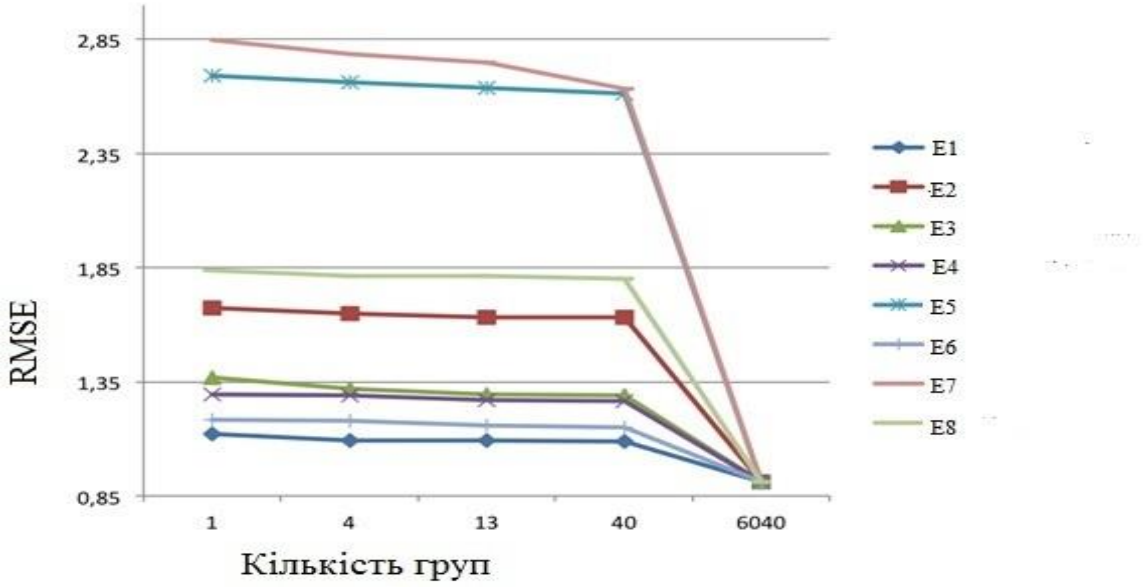


Рис. 4.17 – Залежності похибки прогнозування рекомендацій RMSE для різних моделей прогнозування в групах

На рис. 4.17 наведені результати експериментального дослідження залежностей похибки прогнозування рекомендацій RMSE для різних моделей прогнозування в групах:

E1 – адитивна утилітарна модель; E2 – модель середнього значення без найменшого задоволення (порогове значення 1); E3 − модель середнього

значення без найменшого задоволення (порогове значення 2); E4 − модель середнього значення без найменшого задоволення (порогове значення 3); E5 − модель середнього значення без найменшого задоволення (порогове значення 4); E6 – модель голосування схваленням; E7 – модель найменшого задоволення; E8 – модель найбільшого задоволення.

Результати експериментальних досліджень показують, що найкращі результати дає адитивна утилітарна модель.

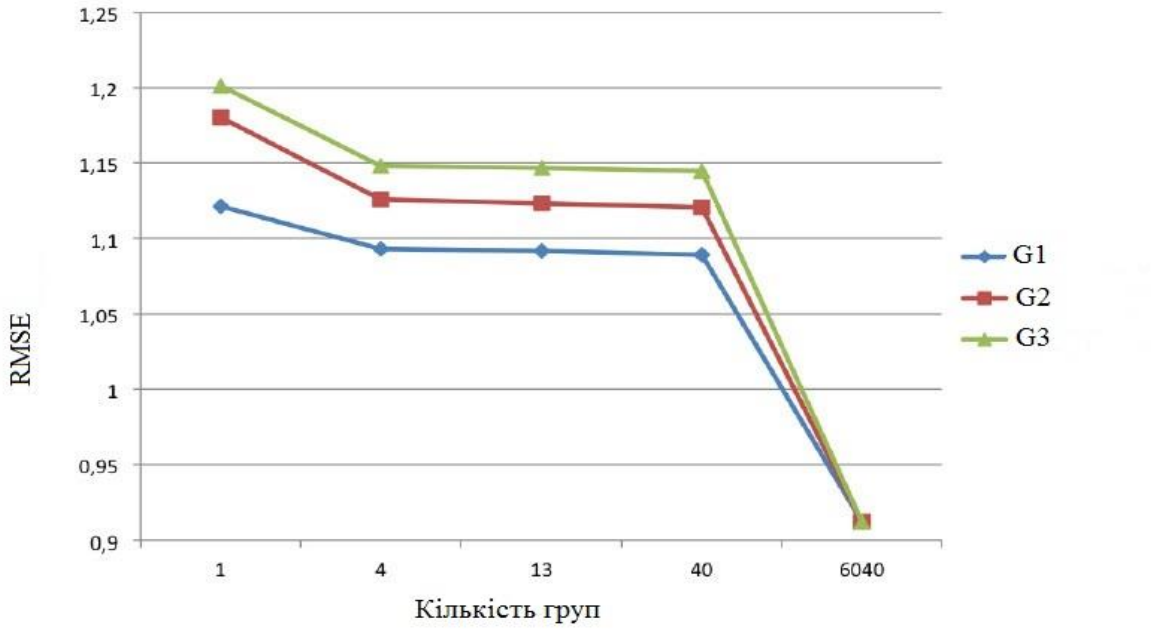


Рис. 4.18 – Залежності похибки прогнозування RMSE для різної кількості оцінених предметів в групі для адитивної утилітарної моделі

На рис. 4.18 приведено залежності похибки прогнозування RMSE для різної кількості спільно оцінених предметів в групі:

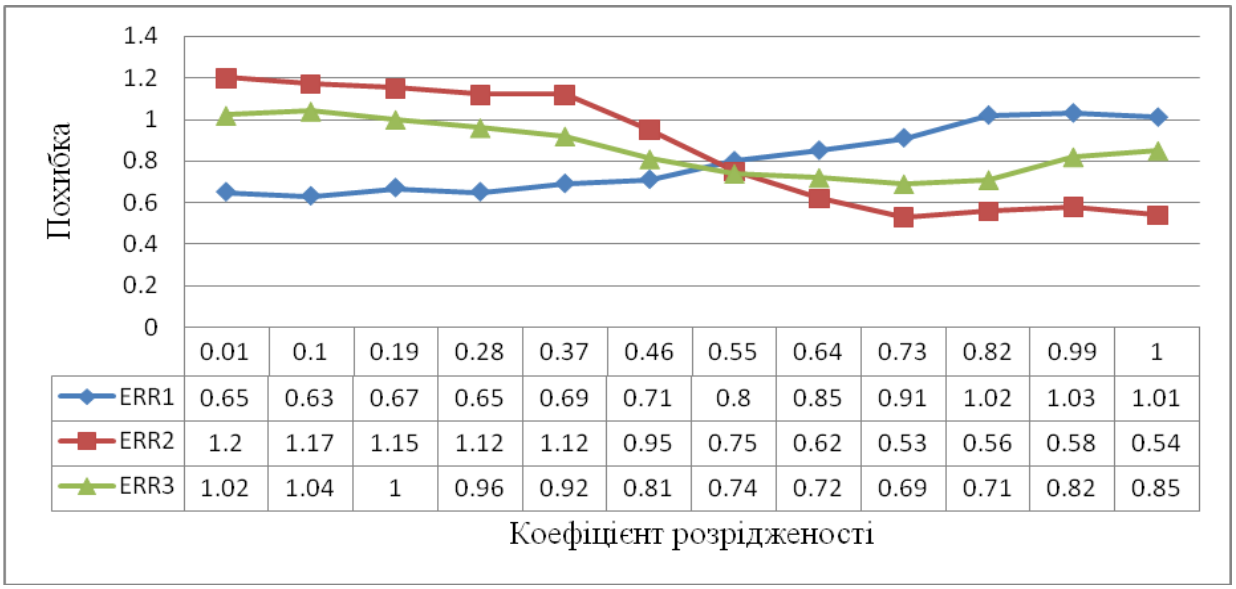
G1 – кількість оцінених предметів в групі 10%;

G2 − кількість оцінених предметів в групі 15%;

G3 − кількість оцінених предметів в групі 20%.

Експериментальні дослідження показують, що чим більша кількість спільно оцінених предметів в групі, тим менша точність прогнозування. Це пов’язано з тим, що із зростанням кількості оцінених предметів в групі

зменшується кількість спільно оцінених предметів в профілях користувачі. Цей ефект зумовлений великою розрідженістю матриці користувач-предмет.



	0.01	0.1	0.19	0.28	0.37	0.46	0.55	0.64	0.73	0.82	0.99	1
ERR1	0.65	0.63	0.67	0.65	0.69	0.71	0.8	0.85	0.91	1.02	1.03	1.01
ERR2	1.2	1.17	1.15	1.12	1.12	0.95	0.75	0.62	0.53	0.56	0.58	0.54
ERR3	1.02	1.04	1	0.96	0.92	0.81	0.74	0.72	0.69	0.71	0.82	0.85

Рис. 4.19 – Залежність похибки прогнозування рекомендацій для тестового набору MoviLens в залежності від коефіцієнту розрідженості матриці користувач-предмет

На рис. 4.19 наведені результати дослідження точності методу прогнозування рекомендацій для груп користувачів в залежності від значення коефіцієнта розрідженості матриці користувач-предмет. Дослідження проводилися для кількості груп – 8.

ERR1 – залежність похибки прогнозування від розрідженості матриці користувач-предмет для методу кластеризації числових профілів користувачів; ERR2 – залежність похибки прогнозування від розрідженості матриці користувач-предмет для методу двохетапної категоріально-числової кластеризації профілів користувачів; ERR3 – залежність похибки прогнозування від розрідженості матриці користувач-предмет для методу мішаної категоріально-числової кластеризації профілів користувачів.

Наведені тестові результати підтверджують ефективність розроблених моделей і методів.

4.10. Висновки до Розділу 4

У **четвертому розділі** дисертаційної роботи розроблено інформаційне забезпечення для тестування моделей і методів прогнозування рекомендацій для інтернет-магазину, розроблена структура математичного забезпечення, розроблена структура програмного забезпечення, яка дозволяє вибрати метод прогнозування рекомендацій, метод пошуку груп користувачів, метод прогнозування рекомендацій в групі користувачів, величину поділу тестової матриці користувач-предмет на прогнозовану і тестову частини, метод розрахунку точності прогнозування рекомендацій. Інформаційне забезпечення розроблене у вигляді реляційної бази даних, яка містить дані із тестового набору даних MoviLens. Наведені результати експериментальних досліджень розроблених моделей, методів і алгоритмів. Експериментальні дослідження проведені на тестовому наборі даних MoviLens. Наведені результати експериментального дослідження коефіцієнтів подібності, які обчислені за косинусною мірою подібності, коефіцієнтом кореляції Пірсона, коефіцієнтом Жаккарда і оберненою евклідовою відстанню, коефіцієнтом подібності, який враховує демографічні характеристики користувачів. Результати експериментальних досліджень показали більшу точність прогнозування рекомендацій при використанні коефіцієнтів подібності, які використовують демографічні характеристики користувачів. Наведені результати експериментальних досліджень залежності точності прогнозування рекомендацій від кількості груп користувачів для різних моделей прогнозування рекомендацій. Показано, що найвищу точність прогнозування дає використання адитивної утилітарної моделі прогнозування рекомендацій. Наведені результати дослідження залежності точності прогнозування рекомендацій від розрідженості матриці користувач предмет для гібридного методу прогнозування рекомендацій.

Основні положення цього розділу викладені у публікаціях автора [1, 5, 7, 9]

ЗАГАЛЬНІ ВИСНОВКИ

У дисертаційній роботі розв’язано актуальне науково-прикладне завдання, яке полягає у розробленні моделей і методів для прогнозування рекомендацій інтернет-магазином для кінцевого користувача. При цьому отримано такі результати:

1. Виконано аналіз актуального стану моделей, методів, засобів і алгоритмів побудови рекомендаційних систем. Це дало можливість визначити переваги і недоліки сучасного математичного забезпечення рекомендаційних систем.
2. Розроблено формальну теоретико-множинну модель колаборативної рекомендаційної системи. Виведена цільова функція, яку необхідно мінімізувати в процесі роботи колаборативної рекомендаційної системи.
3. Розроблено метод розрахунку коефіцієнтів подібності векторів характеристик користувачів і предметів який, на відміну від існуючих, враховує демографічні і контентні характеристики користувачів і предметів і дозволяє розраховувати коефіцієнти подібності для нового користувача. Проведені експериментальні дослідження у порівнянні з існуючими методами розрахунку коефіцієнтів подібності на тестовій БД MoviLens показали збільшення точності прогнозування на 4%.
4. Розроблено метод прогнозування рекомендацій для користувачів інтернет-магазину, який використовує алгоритм пошуку асоціативних правил Apriori. Метод враховує рівень інтересу кожного користувача до відповідного предмета. Отримані рішення підтверджуються апробацією на тестовій БД MoviLens.
5. Вперше на основі концепції застосування в одному методі категоріальної, мішаної і числової кластеризації розроблено метод пошуку груп користувачів, який адаптується до розрідженості матриці користувач-предмет. Отримані рішення підтверджуються апробацією на тестовій БД MoviLens.

6. Отримав подальший розвиток метод кластеризації мішаних категоріально-числових даних, який, на відміну від існуючих, дозволяє автоматично визначати центри кластерів і використовується при виділенні груп користувачів за мішаними рейтинго-демографічними векторами профілів користувачів. Проведені експериментальні дослідження розробленого методу на тестовій БД MoviLens показали збільшення точності прогнозування рекомендацій на 5% у порівнянні з застосуванням методів чіткої числової кластеризації.

7. За допомогою поділу тестової матриці користувач-предмет MoviLens на тестову і прогнозовану частини проведено тестування точності методів і алгоритмів прогнозування рекомендацій.

8. Розроблено програмне забезпечення, яке включає методи прогнозування рекомендацій, методи розрахунку подібностей векторів профілів користувачів і предметів, методи тестування точності.

9. Результати дисертаційної роботи впроваджені в навчальний процес.

СПИСОК ВИКОРИТАНИХ ДЖЕРЕЛ

1. Лобур М. Моделі і методи прогнозування рекомендацій для колаборативних рекомендаційних систем / М.Лобур, М.Шварц, Ю.Стех // Вісник Національного Університету «Львівська політехніка». Інформаційні системи та мережі, Львів. – 2018. − № 901. − С. 68−75.
2. Stekh Y. Some methods for improving the accuracy of prediction recommendations / Y.Stekh, M.Lobur, M.Shvarts // Вісник Націнального Університету «Львівська політехніка». Комп'ютерні системи проектування. Теорія і практика. Львів. − 2017. − № 882. − С. 46−49.
3. Лобур М. Метод і алгоритм прогнозування рекомендацій для спільнот користувачів / М.Лобур, Ю.Стех., М.Шварц. // Збірник наукових праць Української Академії Друкарства. Кваліlогія книги. Львів, 2017. − № 1 (31). − С. 88−93.
4. Лобур М. Побудова асоціативних правил для прогнозування рекомендацій в колаборативних рекомендаційних системах / М.Лобур, Ю.Стех, М.Шварц // Збірник наукових праць Української Академії Друкарства. Кваліlогія книги. Львів. – 2017. − № 2 (32). − С. 82−86.
5. Lobur M. Application of Recommender Systems in the Design of Complex Microsystem Devices / M.Lobur, M.Shvarts, Y.Stekh // International Journal of Advanced Research in Computer Engineering & Technology. − 2018. − V. 7. − № 9. − P. 709−714.
6. Shvarts M. Analysis of the Effectiveness of Similarity Measures for Recommendations Systems / M.Shvarts, M.Lobur, Y.Stekh. – In: The Experience of Design and Application of CAD Systems in Microelectronics: Proc. of the 14th International Conference, Polyana-Svalyava (Zakarpattya), 21-25 February. Lviv, 2017. − P. 275−277.
7. Shvarts M. Some Trends in Modern Recommender Systems / M.Shvarts, M.Lobur, Y.Stekh – In: Perspective technologies and methods in MEMS

design: Proc. of the 13th International Conference, Polyana-Svalyava (Zakarpattya), 20-23 April. 2017, Lviv. − P. 167−169.

8. Shvarts M. Some Methods for Predicting Recommendations for MEMS Designer Communities / M.Shvarts, M.Lobur, Y.Stekh, I.Demkiv – In: Perspective technologies and methods in MEMS design: Proc. of the 14th International Conference, Polyana-Svalyava (Zakarpattya), 18-22 April, 2018, Lviv. − P. 196−199.
9. Шварц М. Моделі і методи побудови рекомендаційних систем / М.Шварц, Ю.Стех. − Проблеми та перспективи розвитку економіки і підприємництва та комп'ютерних технологій в Україні: зб. тез XIII науково-практична конференції, м.Львів, 2017, Львів. − С. 37−38.
10. Лобур М. Метод прогнозування рекомендацій з врахуванням інтересу спільноти користувачів / М.Лобур, Ю.Стех, М.Шварц. – Комп'ютерне моделювання та програмне забезпечення інформаційних систем і технологій: зб. тез третьої Всеукраїнської науково-практичної конференції м.Рівне, 29-30 вересня 2017, Рівне. − С. 135−137.
11. Лобур М. Використання демографічних характеристик користувачів при прогнозуванні рекомендацій / М.Лобур, Ю.Стех, М.Шварц. − Комп'ютерне моделювання та програмне забезпечення інформаційних систем і технологій:. зб. тез третьої Всеукраїнської науково-практичної конференції м.Рівне, 29-30 вересня 2017, Рівне. − С. 138−139.
12. Lobur M. The method of sequential clustering for predicting recommendations / M.Lobur, M.Shvarts, Y.Stekh – In: CAD in Machinery Design-Implementation and Education Problems: Proc. of the XXV Polish-Ukrainian conference: Bielsko Biala, 20-21 October, Bielsko Biala, 2017. − P. 19−20.
13. Lobur M. The Method and Algorithm for Increasing Diversity in Recommendation Systems / M.Lobur, M.Shvarts, Y.Stekh − In: CAD in Machinery Design-Implementation and Education Problems Issues: Proc. of

the XXVI th International Ukrainian-Polish Scientific and Technical Conference, Lviv, 2018. − P. 110−114.

14. Kosobutskyy P. Geometric calculation of Pi using the Monte Carlo method / P.Kosobutskyy, A.Kovalchuk, M.Kuzmynykh, M.Shvarts – In: Perspective technologies and methods in MEMS design: Proc. of the 12th International Conference, Polyana-Svalyava (Zakarpattya), 20-24 April 2016. Lviv, 2016. − P. 167–169.
15. Закон україни "Про електронну комерцію" / Відомості Верховної Ради (ВВР).− 2015.− № 45.− 410 с.
16. Плескач В.Л. Електронна комерція: [підручник] / В.Л. Плескач, Т.Г. Затонацька.– К.: Знання, 2007. – 535 с.
17. Плескач В.Л. Технології електронного бізнесу.– К.: КНТЕУ, 2004. – 222 с.
18. Шердани А. Анализ экономической эффективности интернет магазинов. Критерий Шердани. / ООО «Издательский дом Гребенников»: Интернет маркетинг.− 2008.− № 02(44). − С. 98−109
19. Шалева О.І. Електронна комерція: Навчальний посібник / О.І. Шалева — К.:Центр учбової літератури, 2011. − 209 с.
20. Тардаскіна Т.М. Електронна комерція: Навчальний посібник / Т.М. Тардаскіна, Є.М. Стрельчук, Ю.В. Терешко. – Одеса: ОНАЗ ім. О.С. Попова, 2011. – 244 с.
21. Huang Z. A graph model for ecommerce recommender systems / Z. Huang, W. Chung, H. Chen // Journal of the American Society for Information Science and Technology. – 2004. – V. 55. – № 3. – P. 259–274.
22. Linden G. Amazon.com recommendations: item-to-item collaborative filtering / G. Linden, B. Smith, and J. York // IEEE Internet Computing. – 2003. – V. 7. – № 1. − P. 76–80.
23. Resnick P. GroupLens:An Open Architecture for Collaborative Filtering of Netnews / P. Resnick // Proceedings of CSCW ’94, Chapel Hill, NC, 1994. – P. 571–585.

24. Резникова Н.П. Маркетинг в телекоммуникациях / Н.П.Резникова – М.: Эко-Трендз, 2002. – 336 с.

25. Балабанов И.Т. Электронная коммерция / И.Т. Балабанов – СПБ: Питер, 2001. – 336 с.

26. Юрасов А.В. Электронная коммерция: Учебное пособие / А.В. Юрасов – М.: Дело, 2003. – 480 с.

27. Интернент-магазин: организация, налогообложение, учет / Ред. Кавторева Я. – Харьков: Фактор, 2009. – 128 с.

28. Холмогоров В. Интернет-маркетинг. Краткий курс. 2-е издание / В. Холмогоров – СПб.: Питер, 2002. – 272 с.

29. Сальников С. Успешный интернет-магазин с нуля / С. Сальников – Интернет издание, 2014. – 125 с.

30. Соловьев Д., Писарев А. Интернет-магазин без правил / Д.Соловьев, А.Писарев – СПб: Питер, 2014. – 162 с.

31. Орлов Л.В. Как создать электронный магазин в Интернет. 2-е изд. / Л.В.Орлов – М.: Бук-пресс, 2006. – 384 с.

32. Бабаев А.Б.Создание сайтов / А.Б.Бабаев, Н.В.Евдокимов, М.М.Боде − СПб.: Питер, 2014. – 410 с.

33. Королёв Р.Взлом конверсии. Как сделать сайт, который будет продавать / Р.Королёв − Издательские решения, 2018. − 110 с.

34. Кругляк Ю.А. Web-программирование и Web-дизайн / Ю.А.Кругляк, Л.С.Кострицкая – Экология, 2011.− 156 с.

35. Еремеевский А. Интернет-магазин с нуля. Полное пошаговое руководство / А. Еремеевский, К. Акила – Изд-во:Питер, 2013.− 120 с.

36. Salton G. Automatic text processing: The transformation, analysis, and retrieval information by computer / G. Salton // Addison Wesley, 1989. – 530 p.

37. Wierenga B. Handbook of Marketing Decision Models / B. Wierenga − Publisher: Springer US, 2008.− 630 p.

38. Principles of forecasting: a handbook for researchers and practitioners / J.S. Armstrong, ed. − Springer Science & Business Media, 2001.− 350 p.

39. Murthi S. The Role of the Management Science in Research on Personalization / S. Murthi, S. Sarka // Management Science. – V. 49. – N.10. – 2003. − P. 1344−1362.

40. Christakopoulou E. Local Item-Item models for top-N Recommendation / E. Christakopoulou, G. Karypis. – In: Proc. of the 10th ACM Conference on Recommender Systems (RecSys 2016), 2016. − P. 67−74.

41. Adomavicius G. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions / G.Adomavicius, A.Tuzhilin // IEEE transactions on knowledge and data engineering. – V. 17. – N.6. – 2005 − P. 734−749.

42. Blanco-Fernandez Y. Providing entertainment by content-based filtering and semantic reasoning in intelligent recommender systems / Y.Blanco-Fernandez, J.J. Pazos-Arias, A. Gil-Solla, M. Ramos-Cabrer, M. Lopez-Nores // IEEE Transactions on Consumer Electronics. −2008.− V. 54. – N.2. – P. 727–735.

43. Degemmis M. A content-collaborative recommender that exploits wordnet-based user profiles for neighborhood formation / M.Degemmis, P.Lops, G.Semeraro // User Modeling and User-Adapted Interaction. – 2007. – V. 17. – N.3. − P. 217–255.

44. Degemmis M. Integrating Tags in a Semantic Content-based Recommender / M.Degemmis, P.Lops, G.Semeraro, P.Basile // Proc. of the 2008 ACM Conference on Recommender Systems, RecSys 2008. – 2008. − P. 163–170.

45. Di Noia T. Linked Open Data to Support Content-based Recommender Systems / T. Di Noia, R.Mirizzi, V.C.Ostuni, D.Romito, M.Zanker // SEMANTICS 2012 - 8th International Conference on Semantic Systems. − 2012. − P. 1−8.

46. Balaji A. Performance Comparison of Apache Spark and Hadoop Based Large Scale Content Based Recommender System / A.Balaji, A.Sajith // Intelligent Systems Technologies and Applications. ISTA 2017. Advances in Intelligent Systems and Computing. – V. 683. – 2017. − P. 66−73.

47. Magnini B., Strapparava C. Improving User Modelling with Content-based Techniques / B.Magnini, C.Strapparava // Lecture Notes in Computer Science. − V. 2109. – 2001. −P. 74–83.

48. Gemmis L. Content-based recommender systems: State of the art and trends / L.Gemmis, G. Semeraro − In: Recommender system handbook. – 2011. − P. 73−105.

49. Ning X. SLIM: Sparse linear methods for top-n recommender systems / X. Ning, G. Karypis. – In: Data Mining (ICDM), 2011 IEEE 11th International Conference. – 2011. − P. 497−506.

50. Deshpande M. Item-based top-n recommendation algorithm / M. Deshpande,G. Karypis // ACM Transaction on Information Systems − V. 22 − N.1. − P. 143−177.

51. Pazzani M.A. Framework for Collaborative, Content Based and Demographic Filtering / M.A. Pazzani // Artificial Intelligence Review. – 1999. − P. 393−408.

52. Yahya M. User profiling approaches for demographic recommender systems / M. Yahya // Knowledge-Based Systems. − V. 100. − N.15 – 2016. – P. 175−180.

53. Krulwich B. Lifestyle Finder: Intelligent User Profiling Using Large-Scale Demographic Data / B. Krulwich // Artificial Intelligence Magazine. – 1997. – V. 18 – N.2. − P. 37−45.

54. Vozalis M. Collaborative filtering enhanced by demographic correlation / M. Vozalis, K. G. Margaritis. − In: Proceedings of AIAI Symposium on Professional Practice in AI, of the 18th World ComputerCongress, 2004. – P. 105−110.

55. Vozalis M. Using SVD and demographic data for the enhancement of generalized collaborative filtering / M. Vozalis, K. Margaritis // Information Sciences. – 2007. – V. 177. – N.15. – P. 3017–3037.

56. Safoury L. Exploiting User Demographic Attributes for Solving Cold-Start Problem in Recommender System / L. Safoury. A. Salah // Lecture Notes on Software Engineering. – 2013. – V. 1. – N.3. − P. 215−231.

57. Bobadilla J. Recommender systems survey / J. Bobadilla, F. Ortega, A. Hernando, A. Gutirrez // Knowledge-Based Systems. – 2013. – V. 26. – P. 109–132.

58. Felfernig A. Constraint-based recommender systems: technologies and research issues / A.Felfernig, R.Burke. − In: Proceedings of the 10th International Conference on Electronic Commerce, ICEC'08, 2008. − P. 1–10.

59. Felfernig A. An integrated environment for the development of knowledge-based recommender applications / A.Felfernig, R.Burke // International Journal of Electronic Commerce. – 2007. – V. 11. –N.2 – P. 11–34.

60. Bridge D. Case-based recommender systems / D.Bridge, M.Goeker, L.McGinty, B.Smyth // Knowledge Engineering Review. – 2005. – V. 20. – N.3. – P. 315–320.

61. Burke R. Knowledge-Based Recommender Systems / R. Burke // Encyclopedia of Library and Information Science. – 2000. – V. 69. – N.32. – P. 180–200.

62. Lorenzi F. Case-based recommender systems: A unifying view / F.Lorenzi, F.Ricci, R.Tostes, R.Brasil // Lecture Notes in Computer Science. – 2005. – N. 3169 − P. 89–113.

63. Felfernig A. Consistency-based diagnosis of configurationknowledge bases / A.Felfernig, G.Friedrich, D.Jannach, M.Stumptner // Artificial Intelligence. − 2004.– V.152. – N.2. − P. 213–234.

64. Felfernig A. An integrated environment for the developmentof knowledge-based recommender applications. / A.Felfernig, G.Friedrich,

D.Jannach, M.Zanker // International Journal of ElectronicCommerce – 2007. − V. 11. − N.2. – P. 11–34.

65. Peischl B. Recommending effort estimation methods for software project management / B.Peischl, M.Nica, M.Zanker, W.Schmid − In: Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology − WPRRS Workshop, 2009 − V. 3. − P. 77–80.

66. Jameson A. More than the sum of its members: challenges for group recommender systems / A.Jameson − In Proceedings of the working conference on Advanced visual interfaces, 2004 – P. 48–54.

67. Masthoff J. Group Modeling: Selecting a Sequence of Television Items to Suit a Group of Viewers / J.Masthoff // User Modeling and User-Adapted Interaction −2004. – V. 14. – N.1 – P. 37–85.

68. Masthoff J. In pursuit of satisfaction and the prevention of embarrassment: affective state in group recommender systems / J. Masthoff, A. Gatt // User Modeling and User-Adapted Interaction. − 2006.− V. 16. – P. 281–319.

69. McCarthy K. The needs of the many: a case-based group recommender system. / K. McCarthy, L. McGinty, B. Smyth, M. Salamo // Advances in Case-Based Reasoning. – 2006. – P. 196–210.

70. Ali I. Group recommendations: approaches and evaluation / I. Ali, S. W. Kim – In: Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication. ACM. − 2015. – P. 105–110.

71. Christensen I. Social influence in group recommender systems / I. Christensen, S. Schiaffino // Online Information Review – 2014. – V. 38. − N.4 – P. 524–542.

72. Baltrunas L., Makcinskas T., Ricci F. Group recommendations with rank aggregation and collaborative filtering / L.Baltrunas, T.Makcinskas, F.Ricci – In: Proceedings of the fourth ACM conference on Recommender systems, RecSys'10. Barcelona, Spain. − 2010. – P. 119–126.

73. Gartrell M. Enhancing group recommendation by incorporating social relationship interactions / M. Gartrell, X. Xing, Q. Lv, A. Beach, R. Han, S. Mishra, K. Seada. – In: Proceedings of the 16th ACM international conference on supporting group work. ACM. − 2010.− P. 106−110.

74. Guzzi F. Interactive multi-party critiquing for group recommendation. / F. Guzzi, F. Ricci, R. Burke. – In: Proceedings of the 5th ACM Conference on Recommender systems. − 2011. – P. 265–268.

75. Ahn H.J.A. New similarity measure for collaborative filtering to alleviate the new user cold-starting problem / H.J.Ahn // Information Sciences. – 2008. – V. 178. – P. 37–51.

76. Bobadilla J. A collaborative filtering approach to mitigate the new user cold start problem / J.Bobadilla, F.Ortega, A.Hernando, J.Bernal // Knowledge Based Systems. – 2012. − V. 26. – P. 225–238.

77. Heung-Nam K. Collaborative error-reflected models for cold-start recommender systems / K.Heung-Nam, E.S.Abdulmotaleb, J.Geun-Sik // Decision Support Systems. – 2011. − V. 51 –N.3. – P. 519–531.

78. Lam X.N. Addressing cold-start problem in recommendation systems / X.N.Lam, T.Vu, T.D.Le, A.D.Duong −In: Conference On Ubiquitous Information Management And Communication, 2008. − P. 208–211.

79. Loh S. Identifying similar users by their scientific publications to reduce cold start in recommender systems / S.Loh, F.Lorenzi, R.Granada, D.Lichtnow, L.K.Wives, J.P.Oliveira. − In: Proceedings of the 5th International Conference on Web Information Systems and Technologies (WEBIST2009), 2009. − P. 593–600.

80. Martinez L. Incomplete preference relations to smooth out the cold-start in collaborative recommender systems / L.Martinez, L.G.Perez, M.J.Barranco. − In: Proceedings of the 28th North American Fuzzy Information Processing Society Annual Conference (NAFIPS2009), 2009. − P. 1–6.

81. Schein A.I. Methods and metrics for cold-start recommendations / A.I.Schein, A.Popescul, L.H.Ungar, D.M.Pennock. − In: Proceeding SIGIR ’02 Proceedings ofthe 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2002. − P. 253–260.

82. Zhang Y. Some challenges for context-aware recommender systems / Y.Zhang, L.Wang. − In: 5th International Conference on Computer Science and Education (ICCSE), 2010. − P. 362–365.

83. Adomavicius G. Incorporating contextual information in recommender systems using a multidimensional approach / G.Adomavicius, R.Sankaranarayanan, S.Sen, A.Tuzhilin // ACM Transactions on Information Systems – 2005. – V. 23. – P. 103–145.

84. Melville P. Content-boosted collaborative filtering for improved recommendations / P.Melville, R.J.Mooney, R.Nagarajan. − In: AAAI/IAAI, 2002. − P. 187–192.

85. Billsus D. Learning collaborative information filters / D.Billsus, M.J.Pazzani − In: ICML, 1998. − P. 46–54.

86. Malekzadeh E. Recommending the long tail items through personalized diversification / E. Malekzadeh, H. MarjanKaedi // Knowledge-Based Systems. – 2019 − V. 164. − N.15. – P. 348−357.

87. Salton G. Introduction to Modern Information Retrieval / G. Salton, M.J. McGill. − New York, McGraw- Hill, Inc. – 1986. − P. 251−315.

88. Sarwar B. Item-based collaborative filtering recommendation algorithms / B. Sarwar, G. Karypis, J. Konstan, J. Riedl. −In: Proceedings of the 10th International Conference on World Wide Web, 2001. − P. 285–295.

89. Ekstrand M.D. Collaborative filtering recommender systems / M.D. Ekstrand, J.T. Riedl, J.A. Konstan // Found. Trends Human–Comput. Interact. – 2011. –V. 4. –N. 2.− P. 81–173.

90. Деза Е.И. Энциклопедический словарь расстояний / Е.И. Деза, М.М. Деза − М.: «Наука», 2008. – 444 с.

91. Shardanand U. Social information filtering: algorithms for automating ‘‘Word of Mouth’’ / U. Shardanand, P. Maes. − In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1995. − P. 210–217.

92. Jaccard P. The Distribution of the flora in the alpine zone / P. Jaccard // New Phytologist. − 1912. – V. 11. – P. 37–50.

93. Bobadilla J. A new collaborative filtering metric that improves the behavior of recommender systems / J. Bobadilla, F. Serradilla, J. Bernal // Knowl.-Based Syst. −2010. – V. 23 − N6. – P. 520–528.

94. Egghe L. Introduction to Informetrics : quantitative methods in library, documentation and information science / L. Egghe, R. Rousseau. – 1990, Elsevier Science Publishers. – 421 p.

95. Bachrach Y. Sketching for Big Data Recommender Systems Using Fast Pseudo-random Fingerprints / Y.Bachrach, E.Porat // Lecture Notes in Computer Science. – 2013. – V.7966. – P. 459-471.

96. Chen T.T. Research Paper Recommender Systems on Big Scholarly Data / T.T.Chen, M.Lee // Lecture Notes in Computer Science. – 2018. − V. 11016. – P. 251−260.

97. Bellandi V. Designing a Recommender System for Touristic Activities in a Big Data as a Service Platform / V.Bellandi, P.Ceravolo, E.Damiani, E.Tacchini.− In: Innovations in Big Data Mining and Embedded Knowledge. Intelligent Systems Reference Library – 2019. – V. 159. – P. 13−33.

98. Yousfi S. Mixed-Profiling Recommender Systems for Big Data Environment / S.Yousfi, M.Rhanoui M., D.Chiadmi. − In: Lecture Notes in Real-Time Intelligent Systems. RTIS 2017. Advances in Intelligent Systems and Computing. – 2019. – V. 756. – P. 79−89.

99. Singh I. Big Data Analytics Based Recommender System for Value Added Services (VAS) / I.Singh, K.V.Singh, S.Singh. − In: Proceedings of Sixth International Conference on Soft Computing for Problem Solving. Advances in Intelligent Systems and Computing – 2017. – V. 547. – P. 142−150.

100. Baklanov M.A. Linear TV Recommender Through Big Data / M.A.Baklanov, O.E.Baklanova. − In: Data Mining and Big Data. DMBD 2016. Lecture Notes in Computer Science Springer, Cham. – 2016. – V. 9714. – P. 466−474.

101. Li B. Exploring GTRS Based Recommender Systems with Users of Different Rating Patterns / B.Li, J.Yao.− In: Rough Sets. IJCRS 2018. Lecture Notes in Computer Science. Springer, Cham – 2018. – V. 11103. – P. 405−417.

102. Aggarwal C. Data Mining / C.Aggarwal – Switzerland: Springer International Publishing, 2015. – 727 p.

103. Gorunescu F. Data Mining Concepts, Models and Techniques / F. Gorunescu − 2011 Springer-Verlag, Berlin, Heidelberg, 2011. – 353 p.

104. Biemann C. Text Mining From Ontology Learning to Automated Text Processing Applications / B.Chris, M.Alexander. − Springer International Publishing, Switzerland, 2014. – 221 p.

105. I-Hsien T. Web Mining Applications in E-commerce and E-services / T.I-Hsien,Wu Hui-Ju. − Springer-Verlag, Berlin, Heidelberg, 2009 − 181 p.

106. Wang T. The Ontology Recommendation System in E-Commerce Based on Data Mining and Web Mining Technology / T.Wang − In: Jin D., Lin S. (eds) Advances in Electronic Commerce, Web Application and Communication. Advances in Intelligent and Soft Computing, Springer, Berlin, Heidelberg, 2012 − V. 149. – P. 190−195.

107. Khanian M. Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data / M. K. Najafabadi, M. N. Mahrin, S. Chuprat, H. Sarkan // Computers in Human Behavior – 2015 − V. 67. − P. 113−128.

108. Agrawal R. Mining Associations between Sets of Items in Massive Databases / R. Agrawal, T. Imielinski, A. Swami. – In: Proc. of the 1993 ACM-SIGMOD Int'l Conf. on Management of Data, 1993.− P. 207−216.

109. Agrawal R. Fast Discovery of Association Rules / R.Agrawal, R.Srikant – In: Proc. of the 20th International Conference on VLDB, Santiago, Chile, September 1994. – P. 307−328.

110. Guha S. Rock: A robust clustering algorithm for categorical attributes / Guha S., Rastogi R., Shim K. // Information Systems. −2000 − V. 25. − N. 5. – P. 345−366.

111. Masthoff J. Group Modeling: Selecting a Sequence of Television Items to Suit a Group of Viewers / J. Masthoff // User Modeling and User-Adapted Interaction. − 2004. – V. 14 – P. 37−85.

112. Kelly J. S. Social Choice Theory. An Introduction / J. S. Kelly − Springer-Verlag, Berlin, Heidelberg, 1988, −162 p.

113. MovieLens data, [Електрон. ресурс]. − Режим доступу: http://www.grouplens.org/.

Додаток А. Список публікацій здобувача за темою дисертації та відомості про апробацію результатів дисертації

Наукові праці, в яких опубліковані основні наукові результати дисертації:

1. Лобур М.В. Моделі і методи прогнозування рекомендацій для колаборативних рекомендаційних систем. / М.В.Лобур, М.Є.Шварц, Ю.В.Стех // Вісник Національного Університету Львівська політехніка. Інформаційні системи та мережі. Львів. – 2018. − № 901. − С.68−75.

2. Stekh Y., Lobur M., Shvarts M. Some methods for improving the accuracy of prediction recommendations / Y.Stekh, M.Lobur, M.Shvarts // Вісник Націнального Університету Львівська політехніка. Комп'ютерні системи проетування. Теорія і практика. Львів, − 2017. − № 882. − С.46−49.

3. Лобур М.В., Стех Ю.В., Шварц М.Є. Метод і алгоритм прогнозування рекомендацій для спільнот користувачів / М.В.Лобур, Ю.В.Стех., М.Є.Шварц. // Збірник наукових праць Української Академії Друкарства. Квалілогія книги. Львів, 2017. − № 1 (31). − С.88−93.

4. Лобур М.В., Стех Ю.В., Шварц М.Є. Побудова асоціативних правил для прогнозування рекомендацій в колаборативних рекомендаційних системах / М.В.Лобур, Ю.В.Стех, М.Є.Шварц // Збірник наукових праць Української академії друкарства. Квалілогія книги. Львів – 2017. − № 2 (32). − С. 82−86.

5. Lobur M. Application of recommender systems in the design of complex microsystem devices / M.Lobur, M.Shvarts, Y.Stekh // International Journal of Advanced Research in Computer Engineering & Technology. − 2018. − V. 7. − № 9. − P. 709−714.

Наукові праці, які засвідчують апробацію матеріалів дисертації:

6. Shvarts M., Lobur M., Stekh Y. Analysis of the Effectiveness of Similarity Measures for Recommendations Systems / M.Shvarts, M.Lobur, Y.Stekh – In.The Experience of Design and Application of CAD Systems in Microelectronics: Proceedings of the 14th International Conference, Polyana-Svalyava (Zakarpattya), 21-25 February. Lviv, 2017. − P. 275−277.

7. Shvarts M., Lobur M., Stekh Y. Some Trends in Modern Recommender Systems / M.Shvarts, M.Lobur, Y.Stekh – In: Perspective technologies and methods in MEMS design: Proc. of the 13th International Conference, Polyana-Svalyava (Zakarpattya), 20-23 April. 2017, Lviv. − P. 167−169.

8. Shvarts M., Lobur M., Stekh Y. Demkiv I. Some Methods for Predicting Recommendations for MEMS Designer Communities / M.Shvarts, M.Lobur, Y.Stekh, I.Demkiv – In: Perspective technologies and methods in MEMS design: Proceedings of the 14th International Conference, Polyana-Svalyava (Zakarpattya), 18-22 April, 2018, Lviv, P. 196−199.

9. Шварц М.Є., Стех Ю.В. Моделі і методи побудови рекомендаційних систем / М.Є.Шварц, Ю.В.Стех. − Проблеми та перспективи розвитку економіки і підприємництва та комп'ютерних технологій в Україні: зб. тез XIII науково-практична конференції, м.Львів, 2017, Львів. − С.37−38.

10. Лобур М., Стех Ю., Шварц М. Метод прогнозування рекомендацій з врахуванням інтересу спільноти користувачів. Комп'ютерне моделювання та програмне забезпечення інформаційних систем і технологій: зб. тез третьої Всеукраїнської науково-практичної конференції м.Рівне, 29-30 вересня 2017, Рівне, С.135−137.

11. Лобур М., Стех Ю., Шварц М. Використання демографічних характеристик користувачів при прогнозуванні рекомендацій. Комп'ютерне моделювання та програмне забезпечення інформаційних систем і

технологій: зб. тез третьої Всеукраїнської науково-практичної конференції м.Рівне, 29-30 вересня 2017, Рівне, С.138−139.

12. Lobur M., Shvarts M., Stekh Y. The method of sequential clustering for predicting recommendations / M.Lobur, M.Shvarts, Y.Stekh − CAD in Machinery Design-Implementation and Education Problems: Proceedings of the XXV Polish-Ukrainian conference: Bielsko Biala, 20-21 October, Bielsko Biala, 2017. − P. 19−20.

13. Lobur M., Shvarts M., Stekh Y. The Method and Algorithm for Increasing Diversity in Recommendation Systems / M.Lobur, M.Shvarts, Y.Stekh − In:CAD in Machinery Design-Implementation and Education Problems Issues: Proceedings of the XXVI th International Ukrainian-Polish Scientific and Technical Conference, Lviv, 2018. − P. 110−114.

14. Kosobutskyy P., Kovalchuk A., Kuzmynykh M., Shvarts M. Geometric calculation of Pi using the Monte Carlo method / P.Kosobutskyy, A.Kovalchuk, M.Kuzmynykh, M.Shvarts – In:Perspective technologies and methods in MEMS design: Proceedings of the 12th International Conference, Polyana-Svalyava (Zakarpattya), 20–24 April, Lviv, 2016. − P. 167–169.

Додаток Б. Акт впровадження результатів дисертації

«З А Т В Е Р Д Ж У Ю»
Проректор з науково-педагогічної
роботи Національного університету
«Львівська політехніка»
_____ О. Р. Давидчак
«___» _____ 2019 р.

А К Т
впровадження у навчальний процес результатів дисертаційної роботи
Шварца Михайла Євгенійовича «Гібридні моделі і методи прогнозування рекомендацій для інтернет-магазину» на здобуття
наукового ступеня кандидата технічних наук

Дисертаційна робота аспіранта Шварца Михайла Євгенійовича «Гібридні моделі і методи прогнозування рекомендацій для інтернет-магазину» виконана на здобуття наукового ступеня кандидата технічних наук за спеціальністю 01.05.03 – «Математичне та програмне забезпечення обчислювальних машин і систем» під керівництвом д.т.н., професора М.В. Лобура. Вона спрямована на розв'язання актуальної науково-технічної проблеми забезпечення точності і достовірності прогнозування персоналізованих рекомендацій для інтернет-магазину.

Розроблені наступні моделі, методи і алгоритми: розрахунку коефіцієнтів подібності мішаних векторів для прогнозування рекомендацій методом зваженої суми, які містять категоріальні (демографічні) і числові компоненти; удосконалено метод мішаної кластеризації, який використовує для кластеризації категоріально-числові багатовимірні векторні значення і дозволяє автоматично вибирати кількість і положення центрів кластерів в категоріально-векторному просторі; розроблено метод прогнозування рекомендацій, який базується на концепції асоціативних правил, враховує інтереси існуючих в системі користувачів, дозволяє надавати рекомендації новому користувачу, пропонувати користувачу супутні предмети; для пошуку асоціативних правил розроблено ітеративний метод збільшення підтримки асоціативних правил, який дозволяє зменшити обчислювальну складність; розроблено метод пошуку груп користувачів на основі застосування в одному методі числової, категоріальної і мішаної кластеризації. Наведені вище методи і алгоритми використовуються в процесі викладання дисципліни: «Інноваційні інформаційні технології» для підготовки магістрів за спеціальністю «Інформаційні технології проектування» і «Методи і системи штучного інтелекту» для підготовки бакалаврів за спеціальністю «Комп'ютерні науки»

Акт впровадження затверджено на засіданні кафедри САПР 13 вересня 2018 року, протокол № 2.

Директор Інституту комп'ютерних наук та інформаційних технологій д. т. н., професор — Медиковський М. О.

Зав. кафедри систем автоматизованого проектування д. т. н., професор — Лобур М. В.