

[6] V. Garousi *et al.*, “AI-powered software testing tools: A systematic review and empirical assessment of their features and limitations,” arXiv.org. Accessed: Oct. 12, 2025. [Online]. Available: <https://arxiv.org/abs/2409.00411v3>

УДК 004.89:004.93

ОЦІНКА ВПЛИВУ РІЗНИХ МЕТРИК НА ПРОДУКТИВНІСТЬ СИСТЕМИ РЕКОМЕНДАЦІЙ ФІЛЬМІВ

Маріупольський Р. Ю., Жебко О. О.

(romario311204@gmail.com, al.zhebko@gmail.com)

Миколаївський національний аграрний університет (Україна)

Досліджено вплив різних метрик на оцінку продуктивності систем рекомендацій фільмів. Проаналізовано основні типи метрик: метрики точності прогнозування рейтингу (RMSE, MAE), метрики релевантності (Precision, Recall, F1-score) та метрику ранжування (NDCG). Показано, що RMSE та MAE ефективно оцінюють точність передбачення числових рейтингів, але не враховують релевантність та порядок рекомендацій. Метрики Precision, Recall та F1-score дозволяють визначити відповідність рекомендованих фільмів інтересам користувача, проте ігнорують позицію елементів у списку. NDCG є найповнішою метрикою, що одночасно враховує релевантність та якість ранжування рекомендацій. Обґрунтовано необхідність комплексного підходу до оцінки систем рекомендацій з використанням різних метрик залежно від конкретних цілей застосування та пріоритетів бізнесу.

Сучасні системи рекомендацій стали невід’ємною частиною цифрових платформ, які пропонують широкий вибір контенту. Основна їх задача – допомогти користувачу швидко знайти контент, що відповідає його смакам і інтересам. Проте для створення дійсно ефективної системи важливо не лише розробити якісний алгоритм, але й правильно оцінити його роботу за допомогою метрик продуктивності. Проблема полягає в тому, що різні метрики розкривають різні аспекти роботи системи, і неправильний вибір може призвести до ситуації, коли модель демонструє високі показники на папері, але не задовольняє реальні потреби користувачів. Наприклад, модель може точно прогнозувати рейтинги, але формувати нецікаві або погано організовані списки рекомендацій.

Метою роботи є аналіз впливу різних метрик на оцінку продуктивності системи рекомендацій фільмів та визначення оптимального підходу до їх вибору. Для досягнення мети вирішувалися наступні завдання: 1) проаналізувати метрики точності прогнозування рейтингу (RMSE та MAE) та визначити їх переваги й обмеження; 2) дослідити метрики релевантності (Precision, Recall, F1-score) та їх роль в оцінці якості рекомендованих списків; 3) вивчити метрику ранжування NDCG та її значення для покращення користувацького досвіду; 4) обґрунтувати необхідність комплексного підходу до оцінки систем рекомендацій.

Вибір метрик – це стратегічний крок, що визначає, які аспекти роботи системи будуть покращуватися. Кожна метрика розкриває окремі характеристики продуктивності, тому важливо розуміти їх сильні і слабкі сторони для комплексної оцінки. Дослідження впливу різних метрик дозволяє створювати системи, які не лише точно прогнозують оцінки, але й формують релевантні, зручно організовані рекомендації, що відповідають реальним потребам користувачів та цілям бізнесу. Наприклад, модель може дуже точно прогнозувати числові рейтинги фільмів, але при цьому формувати список рекомендацій, який користувачі вважають нецікавим або погано організованим. Тому важливо розуміти сильні і слабкі сторони кожної метрики, щоб комплексно оцінювати продуктивність [1].

Метрики точності прогнозування рейтингу: RMSE та MAE. Одними з найпоширеніших метрик для оцінки систем рекомендацій є RMSE (Root Mean Square Error) і MAE (Mean Absolute Error). Обидві метрики вимірюють розбіжність між передбаченими рейтингами, які генерує модель, і реальними оцінками користувачів.

RMSE особливо чутлива до великих помилок, оскільки квадрат різниці між передбаченням і фактичним значенням посилює вплив відхилень. Це означає, що моделі з високим RMSE мають тенденцію робити значні помилки у передбаченні рейтингу. Відповідно, при оптимізації моделі за цією метрикою акцент робиться на мінімізації великих помилок, що іноді може призводити до "консервативних" прогнозів – модель намагається уникати крайніх значень і зменшує варіативність оцінок.

MAE, натомість, рахує середнє абсолютне відхилення і тому менш чутлива до викидів. Вона дає більш "прямий" показник середньої помилки, що часто є зрозумілішим для інтерпретації. Обираючи MAE, можна краще зрозуміти, наскільки в середньому прогноз відхиляється від істинного рейтингу.

Проте обидві ці метрики мають суттєвий недолік для задач систем рекомендацій: вони не враховують, який саме список фільмів отримує користувач і в якому порядку вони представлені. В реальному житті кінцевий користувач не просто дивиться на оцінки, а оцінює, наскільки корисними і зручними є самі рекомендації. Тобто навіть при низьких значеннях RMSE або MAE користувач може бути незадоволений системою, якщо вона, наприклад, пропонує йому фільми, які йому нецікаві або розташовані в незручному порядку.

Метрики релевантності: Precision, Recall, F1-score. Щоб оцінити, наскільки рекомендовані фільми відповідають інтересам користувача, використовують метрики релевантності, зокрема Precision, Recall та F1-score.

Precision відображає відсоток рекомендованих фільмів, які дійсно виявилися релевантними для користувача. Якщо Precision висока, це означає, що система рідко рекомендує «помилкові» або нерелевантні фільми. Висока точність є особливо важливою, коли користувач хоче отримати якісні рекомендації і не хоче витрачати час на перегляд непотрібних пропозицій.

Однак Precision не показує, чи система знайшла всі релевантні фільми, які могли б зацікавити користувача. Для цього існує Recall – метрика, що вимірює, яка частка всіх релевантних фільмів була успішно рекомендована. Високий Recall означає, що система не пропускає важливі для користувача фільми, проте це може призводити до зниження точності, якщо система рекомендує занадто багато фільмів, включно з нерелевантними. (рисунк.1).

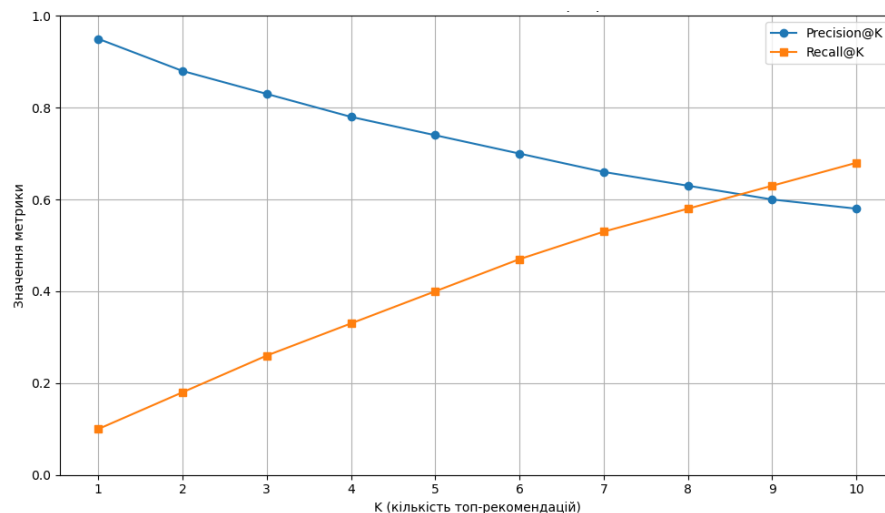


Рис. 1. Precision та Recall в залежності від K (Топ-K рекомендації)

F1-score об'єднує ці дві метрики, представляючи їх гармонійне середнє. Він допомагає знайти баланс між точністю і повнотою рекомендацій, що робить цю метрику зручною для загальної оцінки продуктивності системи.

Попри те, що Precision, Recall і F1-score оцінюють якість рекомендованого списку, вони не враховують порядок, у якому фільми представлені. А оскільки користувачі зазвичай звертають увагу на перші позиції у списку, ігнорування ранжування може знизити якість користувацького досвіду [2].

Метрика ранжування: NDCG. Саме для оцінки якості ранжування у списках рекомендацій широко використовується метрика NDCG (Normalized Discounted Cumulative Gain). Вона поєднує два важливі аспекти: релевантність рекомендованих фільмів та їхню позицію в списку.

Ідея NDCG полягає у тому, що релевантні фільми, розташовані ближче до початку списку, мають більшу вагу для користувача, ніж ті, що знаходяться далеко. Також ця метрика підтримує градації релевантності, тобто різні рівні корисності фільмів для користувача, що робить її дуже гнучкою і точною.

Оптимізація системи за метрикою NDCG стимулює алгоритми не лише знаходити релевантний контент, а й розміщувати його у зручному для сприйняття порядку. Це значно покращує користувацький досвід, адже користувачі швидше знаходять потрібний контент і витрачають менше часу на пошук.

Комплексний підхід до оцінки системи рекомендацій. Кожна з описаних метрик розкриває лише певний аспект роботи системи. Тому для повної оцінки продуктивності доцільно використовувати їх у комплексі.

Наприклад, RMSE і MAE допоможуть оцінити, наскільки точно модель передбачає рейтинги, але щоб зрозуміти, наскільки ці передбачення корисні для кінцевого користувача, необхідно враховувати Precision, Recall, F1-score та NDCG. Відсутність хоча б однієї з цих метрик у аналізі може призвести до неправильних висновків щодо ефективності системи.

Вибір конкретної метрики або їх набору залежить від цілей і контексту застосування системи. Якщо пріоритет – точність передбачення рейтингу, то варто зосередитися на RMSE чи MAE. Якщо головна задача – надати користувачу максимально релевантні рекомендації, краще орієнтуватися на Precision, Recall і F1-score. Якщо ж важливим є зручне ранжування, то основну увагу слід приділити NDCG.

Робота системи рекомендацій фільмів – це складний процес, який потребує багатогранної оцінки. Метрики RMSE і MAE допомагають зрозуміти, наскільки модель точна в прогнозуванні числових оцінок, але не враховують релевантність чи порядок рекомендацій. Метрики Precision, Recall і F1-score відображають, наскільки рекомендовані фільми відповідають інтересам користувача, але не враховують, на яких позиціях вони розміщені. NDCG, в свою чергу, є найповнішою метрикою, що оцінює як релевантність, так і порядок рекомендацій.

У результаті проведеного дослідження вирішено поставлені завдання та отримано наступні результати. По-перше, встановлено, що метрики RMSE та MAE ефективно оцінюють точність прогнозування числових рейтингів, при цьому RMSE є чутливою до великих помилок і стимулює консервативні прогнози, тоді як MAE надає більш зрозумілий показник середньої помилки, проте обидві не враховують релевантність чи порядок рекомендацій. По-друге, доведено, що метрики Precision, Recall та F1-score відображають якість рекомендованого списку та його відповідність інтересам користувача, де Precision показує частку релевантних рекомендацій, Recall – повноту охоплення релевантного контенту, а F1-score забезпечує їх баланс, однак ці метрики не враховують позицію елементів у списку. По-третє, визначено, що NDCG є найповнішою метрикою, яка одночасно оцінює релевантність та якість ранжування, надаючи більшу вагу релевантним фільмам на перших позиціях списку, що значно покращує користувацький досвід. По-четверте, обґрунтовано необхідність комплексного підходу до оцінки систем рекомендацій, де вибір метрик залежить від конкретних цілей: RMSE/MAE – для точності прогнозування, Precision/Recall/F1-score – для релевантності, NDCG – для якості ранжування, що дозволяє створити справді корисну та зручну систему.

Список використаної літератури

- [1] “Математичне та програмне забезпечення системи для рекомендації відеоконтенту”. DSpace :: ELAKPI :: Репозитарій КПІ ім. Ігоря Сікорського. Дата звернення: 22 жовт. 2025. [Онлайн]. Доступно: <https://ela.kpi.ua/handle/123456789/66840>
- [2] М. Аліпа, А. Куриленко та С. Волошук, “Оцінка ефективності рекомендаційних систем IPTV, що враховують неявні уподобання користувачів”, Наука та вир-во, № 23, с. 193–202, січ. 2021. Дата звернення: 22 жовт. 2025. [Онлайн]. Доступно: <https://doi.org/10.31498/2522-9990232020240849>