СЕКЦІЯ XIII.
КОМП'ЮТЕРНА ТА ПРОГРАМНА ІНЖЕНЕРІЯ

DOI 10.62731/mcnd-18.04.2025.002

НЕЙРОМЕРЕЖЕВА ТЕХНОЛОГІЯ ФОРМУВАННЯ РЕКОМЕНДАЦІЙ КОНТЕНТУ ЗА ПРІОРИТЕТАМИ СПОЖИВАЧІВ

Свергун Марк Миколайович
ORCID ID: 0009-0001-0230-4361
аспірант кафедри комп'ютерних наук,
Київський державний торговельно-економічний університет, м. Київ, Україна

Вступ: Сучасні інформаційні системи найчастіше стикаються з проблемою надлишку інформації, що суттєво ускладнює процес пошуку і вибору необхідного контенту користувачами. Обсяг цифрових даних зростає експоненціально, що призводить до ситуації інформаційного перевантаження. У таких умовах особливої ваги набувають рекомендаційні системи, які дозволяють автоматично пропонувати користувачам контент, що відповідає їхнім індивідуальним потребам і пріоритетам [1]. Традиційні підходи до побудови таких систем, а саме : колаборативна (collaborative filtering), та контентна (content-based filtering) фільтрація, що широко використовувалися тривалий час. Вони демонструють суттєві обмеження щодо точності, масштабованості та ефективності. Особливо критичною є проблема «холодного старту» (cold start), яка виникає, коли система не має достатньої інформації про нових користувачів або нові товари. Тому рекомендації виходять неточними або взагалі неможливими, що значно знижує загальну ефективність системи.

Щоб подолати ці обмеження, сучасні дослідження активно використовують нейромережеві підходи, зокрема методи глибинного навчання (deep learning). Такі підходи дозволяють значно покращити точність і персоналізацію рекомендацій, знаходячи приховані закономірності та складні взаємозв'язки в поведінці користувачів. Провідні компанії, такі як Airbnb, Facebook, Google, Netflix, Amazon та

Spotify, успішно впроваджують нейромережеві технології, досягаючи суттєвого зростання залученості аудиторії та покращення користувацького досвіду [2]. Застосування нейронних мереж у рекомендаційних системах мають ключові переваги. Нейронні мережі здатні ефективно працювати з великими масивами даних, і на відміну від класичних лінійних методів, можуть моделювати складні нелінійні взаємозв'язки між користувачем та контентом. Це дозволяє формувати значно точніші прогнози щодо вподобань користувачів. Крім того, нейронні моделі можуть одночасно оптимізувати кілька цілей, наприклад, передбачати як перегляд, так і покупку контенту. Також вони ефективно опрацьовують різнорідні типи даних – текстову інформацію, зображення, відео та контекстні фактори.

Особливу увагу привертають гібридні моделі, які поєднують переваги різних підходів. У данному дослідженні була запропонована нейромережева гібридна система, яка інтегрує колаборативну фільтрацію користувачів і товарів, нейронну колаборативну фільтрацію (Neural Collaborative Filtering, NCF), рекурентні нейронні мережі (RNN) та фільтрацію на основі контенту з використанням алгоритму TF-IDF. Ця система була протестована на наборі даних MovieLens 100k та продемонструвала значне покращення таких метрик, як: зниження RMSE з 0.930 до 0.7723, MAE – з 0.730 до 0.6018, зростання Precision – з 0.730 до 0.8127, Recall – з 0.645 до 0.7312. Крім цього, запропонована модель ефективно вирішує проблему «холодного старту»: для нових користувачів Precision становила 0.762, Recall – 0.685; для нових товарів MAE досягла 0.612, Precision – 0.788, Recall – 0.702 [1]. Такі результати підтверджують перевагу нейромережевих підходів порівняно з класичними алгоритмами, особливо у випадках, коли необхідно враховувати складні або мінливі вподобання користувачів. Інтеграція явних пріоритетів (наприклад, улюблених жанрів або категорій) у нейронні моделі значно підвищує рівень персоналізації рекомендацій, сприяючи позитивному користувацькому досвіду і зростанню бізнес-показників. Таким чином, застосування нейромережевих технологій у формуванні рекомендаційного контенту є ефективним рішенням, яке успішно відповідає сучасним викликам інформаційних систем та потребам цифрового бізнесу [2].

Сучасні рекомендаційні системи на базі нейронних мереж представляють собою складні архітектури, які забезпечують високу

точність і персоналізацію рекомендацій за рахунок використання спільних латентних представлень користувачів та елементів. Основою більшості таких систем є використання шарів ембеддінгів для перетворення категоріальних ознак, таких як ідентифікатори користувачів і контенту, у щільні векторні представлення [2]. Ембеддінги дозволяють моделі виявляти глибокі зв'язки між користувачами та елементами, що є ключовим для створення точних рекомендацій. Тоб то, якщо два користувачі мають схожі історії взаємодій, їхні ембеддінги будуть близькими в латентному просторі, що вказує на схожі вподобання. Найпростіші моделі нейромережевих рекомендацій обчислюють оцінку відповідності користувача та елемента шляхом скалярного добутку їхніх векторних представлень [2]. Цей підхід є аналогічним матричній факторизації, де кожен користувач і елемент описуються прихованими факторами, а їхня взаємодія агрегується лінійно [2]. Однак лінійні моделі мають обмеження у відображенні складних нелінійних взаємодій між ознаками користувачів і контенту. Вони не можуть ефективно моделювати ситуації, коли вподобання користувача залежать від комбінації кількох факторів, таких як жанр фільму, час перегляду та оцінки інших користувачів. Щоб подолати ці обмеження, застосовуються нейронні мережі, які моделюють взаємодії через багатошарові перцептрони (MLP). У таких моделях ембеддінги користувача та елемента об'єднуються, наприклад, шляхом конкатенації чи поелементного множення, і передаються через кілька прихованих шарів з нелінійними активаціями [2]. Ця архітектура, відома як Neural Collaborative Filtering (NCF), дозволяє виявляти приховані залежності між вподобаннями користувачів, що призводить до вищої точності порівняно з традиційними методами колаборативної фільтрації. Дослідження показали, що NCF перевершує методи матричної факторизації, такі як BPR і ALS, на 4–5% за показниками якості рекомендацій [4]. Це досягається завдяки здатності NCF моделювати нелінійні взаємодії, наприклад, враховуючи, як певні комбінації жанрів фільмів можуть впливати на вподобання користувача.

Для врахування послідовності дій користувача широко використовуються рекурентні нейронні мережі (RNN), такі як LSTM і GRU. Ці мережі особливо ефективні в аналізі послідовних патернів у взаємодіях користувачів, що важливо в сценаріях, де порядок дій має значення при рекомендації наступного контенту для перегляду або

покупки [1]. RNNs утримують прихований стан, який накопичує інформацію про попередні взаємодії, дозволяючи моделі розуміти контекст і патерни поведінки користувача протягом часу. Наприклад, якщо користувач переглядає серіали одного жанру протягом тижня, RNN може передбачити, що наступним буде подібний контент. Однак RNNs можуть стикатися з проблемою зникаючих градієнтів при обробці довгих послідовностей, що обмежує їхню ефективність для дуже довгих історій взаємодій, таких як річний перегляд контенту. Сучасні підходи до послідовних рекомендацій все частіше застосовують трансформерні архітектури, які використовують механізм самоуваги (Self-Attention). Трансформери дозволяють одночасно враховувати всю історію взаємодій користувача, вагуючи різні взаємодії залежно від їхньої релевантності до поточного контексту [7]. Це дозволяє моделі вловлювати далекі залежності і покращувати точність прогнозування наступних дій користувача. Наприклад, трансформер може врахувати, що користувач переглядав певний фільм місяць тому, але це все ще впливає на його поточні вподобання. Також в дослідженні показано, що трансформерні моделі значно покращують якість послідовних рекомендацій, особливо при великих обсягах даних, таких як мільйони взаємодій [7].

Інший важливий напрям у розвитку нейромережевих рекомендацій — гібридні системи, які поєднують колаборативні та контентні методи. Нейронні мережі дозволяють інтегрувати контентні характеристики елементів (текстові описи або зображення) безпосередньо у модель. Опис товару може бути перетворений на вектор ознак за допомогою згорткових нейронних мереж (CNN) або трансформерних моделей NLP, а отримані ознаки комбінуються з ембеддінгами користувача [2]. Такий підхід допомагає подолати проблему холодного старту, коли для нових користувачів або елементів відсутня достатня історія взаємодій. Пропонується гібридна система, яка об’єднує традиційну колаборативну фільтрацію, NCF, RNN і контентний аналіз (TF-IDF) [1]. Ця система ефективно долає виклики, такі як розрідженість даних і проблема холодного старту, забезпечуючи високу релевантність рекомендацій. Наприклад, для нового фільму, який ще не має оцінок, система може використовувати опис жанру та акторів, щоб запропонувати його користувачам з подібними вподобаннями. Крім того, нейромережеві рекомендаційні системи часто оптимізують кілька цілей одночасно.

Наприклад, модель може прогнозувати не тільки ймовірність перегляду контенту, але й ймовірність його покупки або підписки [2]. Це досягається шляхом мульти-задачного навчання, де модель навчається на кількох типах взаємодій користувача одночасно. Система може одночасно передбачати, чи користувач перегляне відео, і чи він підпишеться на канал. Також важливо відзначити, що нейронні мережі можуть обробляти різнорідні типи даних, такі як текст, зображення та контекстуальна інформація (час доби або місце користувача), що дозволяє створювати більш комплексні і точні рекомендації [2]. Якщо користувач переглядає рецепти ввечері, система може врахувати цей контекст, щоб запропонувати рецепти для вечері. Отже архітектура та алгоритми нейромережевих рекомендацій постійно розвиваються, пропонуючи все більш просунуті методи для покращення якості і персоналізації рекомендацій. Від простих моделей на основі матричної факторизації до складних гібридних систем, які інтегрують послідовні дані та контентні ознаки, нейронні мережі стали незамінним інструментом у сучасних інформаційних системах. Їхня здатність враховувати пріоритети споживачів і генерувати глибоко персоналізовані рекомендації робить їх особливо цінними для різноманітних застосувань, від e-commerce до соціальних мереж.

Розробка нейромережевої рекомендаційної системи вимагає ефективного опрацювання великих обсягів даних та налаштування складної моделі. На практиці масштабні платформи, такі як YouTube, використовують багаторівневі архітектури для забезпечення високої продуктивності та персоналізації. Наприклад, система рекомендацій YouTube складається з двох етапів: генерації кандидатів та ранжування [3]. На першому етапі глибинна нейронна мережа відбирає невеликий піднабір найбільш релевантних елементів із величезного каталогу, використовуючи модель, натхненну методом "неперервного мішка слів". Ця модель вбудовує розріджені ознаки, такі як ідентифікатори відео, токени пошуку та демографічні дані, у щільні векторні представлення. Для ефективності застосовується вибірковий softmax з тисячами негативних прикладів, що прискорює обчислення в 100 разів порівняно з традиційним softmax. Крім того, використовується ознака "вік прикладу", щоб виправити часозалежну популярність, і модель передбачає майбутні перегляди замість випадкових прихованих переглядів для кращої продуктивності в A/B-тестуванні. На другому етапі інша глибинна

нейронна мережа ранжує ці кандидати, використовуючи багатошарову архітектуру з функцією активації ReLU та враховуючи сотні ознак, таких як минулі взаємодії користувача, історія вражень для прогнозування відтоку та інформація про джерело кандидата. Цей підхід забезпечує як охоплення (recall) рекомендацій, так і точність сортування, залишаючись у межах обчислювальних ресурсів реального часу. Іншим прикладом є модель Deep Learning Recommendation Model (DLRM), розроблена Facebook, яка є передовою моделлю глибокого навчання для завдань персоналізації та рекомендацій [5]. DLRM використовує вбудовування для категоріальних ознак і багатошарову перцептронну мережу (MLP) для обробки неперервних ознак. Вона обчислює другого порядку взаємодії шляхом скалярних добутків між векторами вбудовування та обробленими щільними ознаками, а потім передає їх через верхню MLP для отримання ймовірності взаємодії. Для обробки великих обсягів даних DLRM застосовує комбінацію модельного паралелізму для таблиць вбудовування та паралелізму даних для MLP. Це дозволяє розподілити великі таблиці вбудовування між кількома пристроями та масштабувати обчислення для MLPs. Реалізація DLRM доступна в PyTorch і Caffe2, з відкритим кодом на GitHub, що сприяє її використанню та вдосконаленню.

Розробникам доступні різноманітні фреймворки для побудови таких моделей, зокрема TensorFlow (і бібліотека TensorFlow Recommenders) та PyTorch. Компанія Facebook відкрила реалізацію своєї моделі DLRM з використанням PyTorch, продемонструвавши ефективність паралельного навчання на великих даних [5]. Крім того, NVIDIA пропонує фреймворк HugeCTR, який дозволяє розміщувати вбудовування, більші за пам'ять GPU, на дискових сховищах з незначним впливом на швидкість навчання [2]. Це особливо важливо для рекомендаційних систем, де таблиці вбудовування можуть досягати десятків гігабайт або більше, що перевищує пам'ять одного сервера. Наведений фрагмент коду ілюструє створення простої моделі нейромережевої рекомендації на основі шарів вбудовування користувача і елемента з подальшим щільним шаром:

```
import tensorflow as tf # Приклад нейромережевої моделі рекомендацій
num_users = 10000 # кількість унікальних користувачів
num_items = 5000 # кількість унікальних елементів контенту
latent_dim = 32 # розмірність простору ембеддінгів
```

```
user_input = tf.keras.Input(shape=(1,), name="user_id")
item_input = tf.keras.Input(shape=(1,), name="item_id")
user_emb = tf.keras.layers.Embedding(num_users, latent_dim)(user_input)
item_emb = tf.keras.layers.Embedding(num_items, latent_dim)(item_input)
concat_vec = tf.keras.layers.Concatenate()([tf.reshape(user_emb, [-1, latent_dim]), tf.reshape(item_emb, [-1, latent_dim])])
hidden = tf.keras.layers.Dense(64, activation='relu')(concat_vec)
score_output = tf.keras.layers.Dense(1, activation='sigmoid')(hidden)
model = tf.keras.Model(inputs=[user_input, item_input], outputs=score_output) model.compile(optimizer='adam', loss='binary_crossentropy')
```

Цей код реалізує спрощений підхід Neural Collaborative Filtering (NCF): для кожного користувача та елемента контенту генеруються навчувані латентні вектори-вбудовування, які об'єднуються і пропускаються через невелику багатошарову мережу для прогнозування ймовірності взаємодії (вподобання контенту). У реальних системах модель може мати більше шарів і враховувати додаткові входи (часові ознаки, категорії контенту, соціальні зв’язки тощо) для підвищення точності. Наприклад, можна додати шари для обробки контексту, таких як час доби чи місце розташування користувача, що дозволяє моделі адаптуватися до динамічних умов.

Тренування таких моделей виконується на наборах даних про історію поведінки користувачів; обсяг даних часто вимірюється в мільйонах або мільярдах взаємодій, тому важливим є оптимізація зберігання і читання даних. Ефективне паралельне навчання — ключ до масштабування: великі таблиці вбудовування (містять ваги для всіх користувачів та елементів) можуть сягати десятків гігабайт або більше, що перевищує пам’ять одного сервера [2]. Для вирішення цього завдання використовують розподіл таблиць вбудовування між кількома обчислювальними вузлами та прискорення обчислень на GPU. Фреймворк NVIDIA HugeCTR дозволяє розміщувати вбудовування, більші за пам’ять GPU, на дискових сховищах з незначним впливом на швидкість навчання [2]. Це забезпечує можливість тренування моделей на великих наборах даних без значного збільшення часу обробки. Крім того,

нейронні мережі пропонують значні переваги для систем рекомендацій. Вони можуть досягати вищої точності з більшою кількістю даних, на відміну від традиційних методів машинного навчання, які досягають плато [2]. Нейронні мережі також забезпечують гнучкість у дизайні, дозволяючи навчання з кількома задачами та обробку мультимодальних даних, що робить їх ідеальними для складних рекомендаційних систем. Наприклад, модель може одночасно прогнозувати ймовірність перегляду відео, ймовірність підписки на канал і ймовірність покупки товару, що дозволяє платформам оптимізувати різні бізнес-метрики. Крім того, фреймворки глибокого навчання оптимізовані для обробки великих обсягів даних, що є критично важливим для масштабних рекомендаційних систем [2]. Це включає підтримку розподілених обчислень, автоматичне диференціювання і оптимізацію градієнтного спуску, що значно прискорює процес навчання.

Реалізація нейромережевих рекомендаційних систем вимагає використання спеціалізованих архітектур, таких як двоступеневі системи YouTube або DLRM від Facebook, а також потужних інструментів, таких як TensorFlow, PyTorch і HugeCTR. Ці інструменти дозволяють ефективно обробляти великі обсяги даних і забезпечувати високу точність рекомендацій, що є ключовим для сучасних інформаційних систем. Отже система YouTube може обробляти мільярди переглядів щодня, забезпечуючи персоналізовані рекомендації в реальному часі, а DLRM дозволяє масштабувати моделі для мільярдів користувачів, що робить їх незамінними для великих платформ.

Висновки. Нейромережева технологія формування рекомендаційного контенту є сучасним підходом до персоналізації, який перевершує класичні методи завдяки гнучкості моделювання та здатності вчитися на великих обсягах даних. Глибокі моделі успішно поєднують різні джерела інформації про вподобання користувачів, включаючи їхні явні пріоритети, такі як улюблені жанри чи цінові діапазони, і можуть динамічно адаптувати рекомендації до змін у поведінці. Наприклад, якщо користувач змінює свої інтереси, модель може швидко підлаштуватися, враховуючи нові взаємодії, що робить рекомендації більш релевантними та корисними. Розглянуті алгоритми та архітектурні рішення, такі як Neural Collaborative Filtering (NCF), рекурентні нейронні мережі (LSTM, GRU) та трансформери, належать до галузі комп’ютерної та програмної інженерії. NCF виявляє приховані

залежності між вподобаннями, досягаючи вищої точності порівняно з традиційними методами матричної факторизації, такими як BPR чи ALS, на 4–5% за показниками якості рекомендацій [4]. Рекурентні мережі ефективно аналізують послідовність дій, порядок перегляду контенту, що важливо для прогнозування наступних взаємодій [1]. Трансформери, завдяки механізму самоуваги, дозволяють враховувати всю історію взаємодій, покращуючи точність, особливо для великих наборів даних [7].

Гібридні системи, які поєднують колаборативні та контентні методи, також відіграють важливу роль. Вони інтегрують контентні характеристики, такі як текстові описи чи зображення, безпосередньо в модель, що допомагає подолати проблему холодного старту, коли для нових користувачів чи елементів бракує даних [1]. У дослідженні запропоновано гібридну систему, яка об'єднує традиційну колаборативну фільтрацію, NCF, RNN і контентний аналіз (TF-IDF), забезпечуючи високу релевантність рекомендацій навіть при обмежених даних [1]. Це особливо важливо для платформ, де нові користувачі чи продукти потребують швидкої інтеграції.

Тренування моделей виконується на наборах даних про історію поведінки користувачів, обсяг яких часто вимірюється в мільйонах або мільярдах взаємодій. Для цього важливим є оптимізація зберігання і читання даних, зокрема використання розподілених систем зберігання, таких як Hadoop чи Spark, для швидкого доступу до великих наборів даних. Ефективне паралельне навчання є ключем до масштабування, адже великі таблиці вбудовувань містять ваги для всіх користувачів та елементів, що може перевищувати можливості одного сервера [2]. Для прискорення обчислень часто використовують GPU, що значно зменшує час навчання, особливо для моделей із сотнями мільйонів параметрів.

Подальший розвиток нейромережевих технологій рекомендацій пов'язаний із застосуванням більш потужних моделей, таких як трансформери та великомасштабні мовні моделі, які можуть покращити якість персоналізації. Наприклад, трансформери дозволяють враховувати далекі залежності в історії взаємодій, що особливо корисно для довгих послідовностей [7]. Однак важливим залишається забезпечення довіри до рекомендацій, що включає пояснюваність моделей та відсутність упереджень. Пояснюваність є критичною для збереження довіри користувачів, особливо в контексті використання

великих обсягів персональних даних. Дослідження Zhang et al. (2024) фокусується на пояснюваності глибоких нейронних мереж для рекомендацій, що є ключовим для прозорості моделей. Крім того, етичні аспекти, такі як конфіденційність даних і упередженість моделей, потребують додаткової уваги, особливо в умовах зростання регуляцій, таких як GDPR. Нейромережеві технології рекомендаційного відбору контенту пропонують значний науковий і прикладний потенціал. Вони не лише покращують якість і персоналізацію рекомендацій, але й відкривають нові можливості для досліджень у галузі комп'ютерної та програмної інженерії. Для подальшого просування цієї теми важливо продовжувати дослідження щодо вдосконалення архітектур моделей, оптимізації їхньої продуктивності та забезпечення їхньої прозорості та етичності. Це також включає розробку методів для зменшення енергоспоживання великих моделей, що є важливим у контексті стійкості та екологічності.

Список використаних джерел:

1. Sami A. et al. A deep learning based hybrid recommendation model for internet users // Scientific Reports. – 2024. – Vol.14, Art. 29390 https://surl.li/sgszyd.
2. Schifferer B. Using Neural Networks for Your Recommender System // NVIDIA Technical Blog, https://surl.li/qqoopl.
3. Covington P. *et al.* Deep Neural Networks for YouTube Recommendations // Proc. of ACM RecSys 2016. – 2016. – P. 191–198 https://surl.li/mwppfg.
4. He X. *et al.* Neural Collaborative Filtering // Proc. of WWW '17. – ACM, 2017. https://surl.li/poaeou.
5. Naumov M. *et al.* Deep Learning Recommendation Model for Personalization and Recommendation Systems // *arXiv preprint* arXiv:1906.00091. – 2019 https://surl.li/vyltfd.
6. Cheng H-T. *et al.* Wide & Deep Learning for Recommender Systems // *arXiv preprint* arXiv:1606.07792. – 2016 https://surl.li/mzbtfi.
7. Zivic P., Vazquez H., Sanchez J. Scaling Sequential Recommendation Models with Transformers // Proc. of SIGIR '24. – 2024 https://surl.li/ncdtzo.