

РЕАЛИЗАЦИЯ РЕКОМЕНДАТЕЛЬНОЙ СИСТЕМЫ В МАРКЕТПЛЕЙСЕ НА ОСНОВЕ ЭМБЕДДИНГОВ

Нестеренков С. Н., Лебедев А. В.
Факультет компьютерных систем и сетей,

Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь

E-mail: nesterenkov@bsuir.by, artyomlebedevich@gmail.com

В работе предлагается подход к построению гибридной рекомендательной системы для электронной коммерции на основе векторных представлений. Модель использует компактные 32-мерные эмбединги, объединяющие текстовую информацию, атрибуты товаров и данные о поведении пользователей. Ключевая особенность решения – интеграция контентных векторных сигналов с традиционными методами коллаборативной фильтрации через механизм гибридного переранжирования. Экспериментальная оценка на реальных данных маркетплейса показала увеличение CTR на 6,8%, покрытия каталога на 3% и эффективное решение проблемы холодного старта для новых товаров.

ВВЕДЕНИЕ

Современные маркетплейсы сталкиваются с задачей предоставления персонализированных рекомендаций из каталогов, содержащих десятки и сотни тысяч товарных позиций. Классические методы коллаборативной фильтрации демонстрируют высокую эффективность для популярных товаров, но испытывают серьёзные трудности с новыми позициями (проблема холодного старта) и редкими товарами длинного хвоста [1, 2].

Векторные представления (эмбединги) открывают возможность объединения коллаборативных и контентных подходов: кодирование товаров в едином семантическом пространстве позволяет измерять их близость независимо от истории взаимодействий [3]. Дополнительным преимуществом является возможность применения единой модели не только в рекомендациях, но и в функции поиска [4].

1. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ

Пусть $\mathcal{P} = \{p_1, p_2, \dots, p_N\}$ – множество товаров в каталоге, $\mathcal{U} = \{u_1, u_2, \dots, u_M\}$ – множество пользователей. Целью является построение отображения $f: \mathcal{P} \rightarrow \mathbb{R}^d$, которое кодирует товар p_i в вектор $\mathbf{e}_i \in \mathbb{R}^d$ фиксированной размерности $d = 32$.

Формирование товарных эмбедингов

Представление товара $\mathbf{e}_i$ формируется как взвешенная комбинация признаков:

$$\mathbf{e}_i = W_t \mathbf{t}_i + W_a \mathbf{a}_i + W_b \mathbf{b}_i \quad (1)$$

где $\mathbf{t}_i \in \mathbb{R}^{d_t}$ – текстовые признаки (название, описание, бренд), $\mathbf{a}_i \in \mathbb{R}^{d_a}$ – атрибутивные признаки (категория, цена, характеристики), $\mathbf{b}_i \in \mathbb{R}^{d_b}$ – поведенческие признаки (просмотры, покупки), W_t, W_a, W_b – обучаемые матрицы проекции.

Текстовые признаки извлекаются через предобученный трансформер-энкодер для русского

языка [7]:

$$\mathbf{t}_i = \text{Encoder}(\text{concat}(\text{name}_i, \text{desc}_i, \text{brand}_i))$$

Атрибутивные признаки нормализуются:

$$\mathbf{a}_i = [\text{embed}(\text{cat}_i), \frac{\text{price}_i - \mu_p}{\sigma_p}, \text{attr}_i]$$

где $\text{embed}(\text{cat}_i)$ – эмбединг категории, нормализация цены выполняется по среднему μ_p и стандартному отклонению σ_p в категории.

Поведенческие признаки кодируют популярность:

$$\mathbf{b}_i = [\log(1 + v_i^{7d}), \log(1 + v_i^{30d}), \log(1 + c_i^{90d})]$$

где v_i^{7d}, v_i^{30d} – число просмотров за 7 и 30 дней, c_i^{90d} – число покупок за 90 дней. Логарифмическое преобразование сглаживает влияние экстремально популярных товаров.

Выбор размерности $d = 32$ обусловлен балансом между выразительной способностью и вычислительной эффективностью.

Формирование пользовательских эмбедингов

Представление пользователя строится через взвешенное агрегирование:

$$\mathbf{u}_j = \frac{\sum_{i \in \mathcal{I}_j} w_{ij} \mathbf{e}_i}{\sum_{i \in \mathcal{I}_j} w_{ij}} \quad (2)$$

где $\mathcal{I}_j$ – множество товаров в истории пользователя. Веса учитывают тип взаимодействия и давность:

$$w_{ij} = \alpha_{\text{type}(ij)} \cdot \exp(-\lambda \Delta t_{ij})$$

где α_{type} – базовый вес (покупка: 3.0, корзина: 2.0, просмотр: 1.0), $\lambda = 0.05$ – коэффициент затухания, Δt_{ij} – время с момента взаимодействия в днях.

Для измерения релевантности используется косинусное сходство:

$$s_{\text{embed}}(i, j) = \frac{\mathbf{e}_i \cdot \mathbf{u}_j}{\|\mathbf{e}_i\| \|\mathbf{u}_j\|} \quad (3)$$

Для эффективного поиска применяется индекс FAISS HNSW [5], который обеспечивает приближённый поиск за время $O(\log N)$ вместо линейного перебора.

II. ГИБРИДНАЯ ФУНКЦИЯ СКОРИНГА

Финальный скор объединяет коллаборативные и контентные сигналы:

$$S(i, j) = \alpha_i s_{\text{collab}}(i, j) + (1 - \alpha_i) s_{\text{embed}}(i, j) \quad (4)$$

Коллаборативный скор вычисляется через алгоритм Swing [6]:

$$s_{\text{collab}}(i, k) = \sum_{u, v \in \mathcal{U}} \frac{w_u \cdot w_v}{\alpha_{\text{swing}} + |\mathcal{I}_u \cap \mathcal{I}_v| - 1}$$

где суммирование ведётся по парам пользователей, купивших товары i и k ; w_u, w_v – веса пользователей; $\mathcal{I}_u, \mathcal{I}_v$ – множества купленных товаров; $\alpha_{\text{swing}} = 5$ – параметр сглаживания.

Параметр баланса α_i адаптируется индивидуально:

$$\alpha_i = \begin{cases} 0.2, & \text{если } n_i < 5 \\ 0.5, & \text{если } 5 \leq n_i < 20 \\ 0.75, & \text{если } n_i \geq 20 \end{cases}$$

где n_i – число взаимодействий с товаром за 90 дней. Схема обеспечивает переход от контентного подхода для новых товаров к коллаборативному для зрелых позиций.

III. ОБУЧЕНИЕ ВЕКТОРНОЙ МОДЕЛИ

Модель обучается методом контрастивного обучения на исторических данных взаимодействий. Положительные примеры извлекаются из логов: если пользователь купил или кликнул товар, формируется положительная пара (u_j, p_i) .

Отрицательные примеры формируются тремя стратегиями: случайные негативы из всего каталога (60%), жёсткие негативы из той же категории (30%), внутрипакетные негативы – положительные примеры других пользователей (10%). Соотношение положительных и отрицательных примеров составляет 1:5.

Текстовый энкодер основан на предобученной модели BERT для русского языка с 6 слоями трансформера. Проекционные слои реализованы как двухслойные MLP с активацией GELU и dropout 0.1.

Оптимизация выполняется алгоритмом AdamW с learning rate 3×10^{-4} и weight decay 10^{-2} . Обучение проводится на данных за 6 месяцев в течение 20 эпох с размером батча 512.

IV. ПРИМЕНЕНИЕ В ТОЧКАХ КОНТАКТА

Персональные рекомендации. На главной странице генерируются кандидаты из трёх источников: коллаборативная фильтрация (300 товаров), контентная близость через FAISS (200 товаров), популярные новинки (100 товаров). Финальный топ-20 формируется через переранжирование по формуле (4).

Похожие товары на карточке. Кандидаты извлекаются через Swing-алгоритм (до 500 товаров) и переранжируются с учётом семантической близости. Для новых товаров без истории ($n_i < 3$) используются только эмбединги с поиском в той же категории.

Рекомендации в корзине. Для каждого товара строятся дополнения через граф совместных покупок. Кандидаты ранжируются по релевантности всей корзине через среднее сходство с вектором корзины $\mathbf{e}_{\text{basket}}$.

Персонализация промо. Акционные товары ранжируются по релевантности интересам пользователя, избегая показа нерелевантных скидок.

V. ЭКСПЕРИМЕНТАЛЬНАЯ ОЦЕНКА

Система оценивалась на данных маркетплейса с каталогом 78 тыс. товаров и 850 тыс. активных пользователей за 3 месяца. Проводилось A/B-тестирование на 15% трафика.

Онлайн-метрики

Таблица 1 – Результаты A/B-теста гибридной системы

Метрика	Базовая	Гибридная	Прирост
CTR@10	7,82%	8,35%	+6,8%
Conversion@10	1,64%	1,79%	+9,1%
Add-to-Cart	4,21%	4,53%	+7,6%

Увеличение CTR на 6,8% статистически значимо ($p\text{-value} < 0.01$) и указывает на повышение релевантности. Прирост конверсии на 9,1% напрямую влияет на выручку платформы.

Покрытие и холодный старт

Базовая система обеспечивала попадание в рекомендации 54,2% товаров за месяц. Гибридная система увеличила покрытие до 57,2% (+3,0 п.п.). Длинный хвост (товары с менее 10 взаимодействиями) увеличил представленность с 11,3% до 19,7%.

Для новых товаров (первые 7 дней в каталоге): базовая система показала попадание 14,8%, CTR 2,3%; гибридная – 61,2% и 5,1% соответственно. Рост попадания в 4,1 раза подтверждает эффективность контентного подхода.

Среднее число уникальных категорий в топ-10 увеличилось с 2,8 до 3,4. Индекс концентрации Джини снизился с 0,71 до 0,66, что указывает на более равномерное распределение внимания между товарами.

Анализ параметра α показал оптимум при 0,5–0,7. Адаптивная схема позволяет динамически выбирать оптимальный баланс для каждого товара.

ЗАКЛЮЧЕНИЕ

В работе предложен гибридный подход к построению рекомендательной системы, объединяющий коллаборативную фильтрацию и контентные методы через векторные представления. Использование компактных 32-мерных эмбедингов обеспечивает вычислительную эффективность при сохранении различительной способности.

Ключевые результаты: увеличение CTR на 6,8%, конверсии на 9,1%, покрытия каталога на 3,0 п.п., эффективное решение проблемы холодного старта с четырёхкратным ростом видимости новых товаров. Механизм адаптивного взвешивания позволяет автоматически выбирать оптимальную стратегию в зависимости от доступности статистики [8].

Дальнейшие направления включают интеграцию визуальной модальности через кодирование изображений товаров, разработку специали-

зированных архитектур для электронной коммерции и применение методов активного обучения для адаптации к сезонным трендам.

1. Koren, Y. Matrix Factorization Techniques for Recommender Systems / Y. Koren, R. Bell, C. Volinsky // IEEE Computer. – 2009. – Vol. 42, № 8. – P. 30–37.
2. Huang, J. Embedding-based Retrieval in Facebook Search / J. Huang [et al.] // Proceedings of the 26th ACM SIGKDD Conference. – 2020. – P. 2553–2561.
3. Covington, P. Deep Neural Networks for YouTube Recommendations / P. Covington, J. Adams, E. Sargin // Proceedings of the 10th ACM Conference on Recommender Systems. – 2016. – P. 191–198.
4. Devlin, J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin [et al.] // Proceedings of NAACL-HLT. – 2019. – P. 4171–4186.
5. Malkov, Y. A. Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs / Y. A. Malkov, D. A. Yashunin // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2020. – Vol. 42, № 4. – P. 824–836.
6. Sanz-Cruzado, J. Structural novelty and diversity in link prediction / J. Sanz-Cruzado [et al.] // Proceedings of the 9th International Workshop on News Recommender Systems. – 2018. – P. 12–19.
7. Devlin, J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin [et al.] // Proceedings of NAACL-HLT. – 2019. – P. 4171–4186.
8. Нестеренков, С. Н. Методы построения рекомендательных систем на основе глубокого обучения / С. Н. Нестеренков, А. И. Петров // Доклады БГУ-ИР. – 2023. – Т. 21, № 4. – С. 62–69.