

УДК 004.852:004.822
DOI 10.17513/snt.40899



CC BY 4.0

ДИАГНОСТИЧЕСКИЙ ПРОТОКОЛ ОЦЕНКИ ГРАФОВО-СЕМАНТИЧЕСКИХ РЕКОМЕНДАТЕЛЬНЫХ МОДЕЛЕЙ ПРИ РАЗРЕЖЕННЫХ ДАННЫХ

Дубровин В. В. ORCID ID 0009-0001-0242-366X,
Алаторцев Е. И. ORCID ID 0009-0009-9851-5523,
Бабкин И. А. ORCID ID 0009-0008-1147-8727

*Федеральное государственное бюджетное образовательное учреждение высшего образования
«МИРЭА – Российский технологический университет»,
Москва, Российская Федерация, e-mail: dubrovinor@gmail.com*

В рекомендательных системах с разреженными пользовательско-объектными взаимодействиями агрегированные метрики качества не всегда объясняют, за счет чего получен результат модели. Неоднозначность особенно заметна, когда граф взаимодействий дополняется содержательными признаками объектов: текстовыми описаниями, категориями и метаданными. Цель исследования – разработать компактный диагностический протокол для интерпретации результатов графово-семантических рекомендательных моделей при разреженных данных и проверить его в предметных областях с различной ролью семантических признаков и популярностного смещения. В качестве материалов использованы коллекции Amazon Books, Movies и Yelp, а также подготовленные на их основе экспериментальные срезы. Методика включает полное ранжирование с исключением известных пользователю объектов, семантический и популярностный контроль, стратификацию по популярности объектов и проверку устойчивости ключевых сравнений. Применение протокола позволило различить семантическое доминирование на исходном Amazon, смешанный режим на отфильтрованном срезе Amazon, отрицательный перенос на Movies и популярностное доминирование на Yelp. Полученные результаты показывают, что одинаковое изменение агрегированной метрики может иметь различные объяснения в зависимости от состава данных и базового метода. Протокол уточняет интерпретацию качества моделей в автономных экспериментах, но не заменяет причинную оценку или проверку в пользовательском эксперименте.

Ключевые слова: рекомендательные системы, графовые рекомендательные модели, семантический контроль, стратифицированный анализ, разреженные данные, автономная оценка, популярностное смещение

DIAGNOSTIC EVALUATION PROTOCOL FOR GRAPH-SEMANTIC RECOMMENDER MODELS UNDER SPARSE DATA

Dubrovin V. V. ORCID ID 0009-0001-0242-366X,
Alatortsev E. I. ORCID ID 0009-0009-9851-5523,
Babkin I. A. ORCID ID 0009-0008-1147-8727

*Federal State Budgetary Educational Institution of Higher Education
“MIREA – Russian Technological University”, Moscow, Russian Federation,
e-mail: dubrovinor@gmail.com*

In recommender systems with sparse user-item interactions, aggregate quality metrics do not always explain why a model obtains a particular result. This ambiguity becomes especially visible when interaction graphs are complemented by semantic item features, including textual descriptions, categories, and metadata. The purpose of the study is to develop a compact diagnostic protocol for interpreting graph-semantic recommender models under sparse data and to test it in domains with different roles of semantic features and popularity bias. The study uses Amazon Books, Movies, and Yelp collections and experimental slices prepared from them. The protocol combines full ranking with exclusion of items known to a user, semantic and popularity controls, item-popularity stratification, and stability checks for key comparisons. Its application distinguishes semantic dominance on the original Amazon data, a mixed regime on the filtered Amazon slice, negative transfer on Movies, and popularity dominance on Yelp. The findings show that the same change in an aggregate metric may have different explanations depending on the dataset composition and the reference method. The protocol refines interpretation of recommender-model performance in offline experiments but does not replace causal analysis or evaluation in a user experiment.

Keywords: recommender systems, graph recommender models, semantic control, popularity bias, stratified analysis, sparse data, offline evaluation

Введение

В исследованиях рекомендательных систем новые модели часто сравнивают по агрегированным метрикам ранжирования. Такой подход удобен, но не всегда объясняет природу результата: высокое значе-

ние метрики может быть связано не только с архитектурой модели, но и с особенностями данных, семантической близостью объектов или популярностным смещением.

Для графовых рекомендательных моделей это особенно заметно. LightGCN оста-

ется сильной базовой моделью совместной фильтрации [1], а NCL усиливает графовые представления контрастивным обучением [2]. Современные обзоры фиксируют устойчивый интерес к графовым нейронным сетям в рекомендательных задачах [3]. Однако корректное сравнение таких моделей требует единого разбиения данных, полного ранжирования или корректной выборки негативных объектов и сопоставимых настроек обучения [4, 5]. Дополнительную неопределенность вносит популярностное смещение: оно проявляется уже в простых стратегиях переранжирования [6] и подробно обсуждается в современных обзорах [7, 8]. Содержательные признаки объектов – текст, категории и метаданные – также могут становиться самостоятельным источником высокого качества без полноценного графового обучения. Это видно в классических гибридных подходах [9, 10]. Сходная проблема сохраняется и в современных работах о самоконтролируемом и контрастивном обучении в рекомендательных системах [11, 12].

Причинная и контрфактическая оценка рекомендаций глубже анализирует смещения, но обычно требует информации о политике показа, вероятностях экспозиции или пользовательском эксперименте [13, 14]. В рассматриваемой постановке таких данных нет. Поэтому акцент сделан на автономном диагностическом протоколе, который различает семантическое доминирование, популярностное доминирование, эмпирическое превосходство относительно контролей, смешанный режим и отрицательный перенос.

В российских публикациях последних лет близкие вопросы обсуждаются в работах об архитектурах рекомендательных систем [15] и графовых нейронных сетях в электронной коммерции [16]. Отдельно рассматриваются метрики качества рекомендаций [17], холодный старт, разреженность данных и коррекция релевантности [18, 19]. Для предлагаемого протокола важна эта методическая линия: результат модели рассматривается не как одно число, а как следствие условий эксперимента.

Под диагностическим протоколом далее понимается не новая метрика, а последовательность проверок, сопровождающих основное сравнение моделей. Протокол отделяет случай, когда семантическая ветвь действительно помогает низкосвязным объектам, от ситуации, когда высокое качество уже достигается без обучения графовой модели.

Цель исследования – разработать компактный диагностический протокол для интерпретации результатов графово-семантических рекомендательных моделей при разреженных данных и проверить его на нескольких предметных областях с различной ролью семантических признаков и популярностного смещения.

Материалы и методы исследования

Проверка протокола выполнена на экспериментальных срезах Amazon Books, Movies и Yelp. Исходные данные доступны соответственно по адресам <https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews>, <https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset> и <https://business.yelp.com/data/resources/open-dataset/>. Amazon Books используется как каталог с текстово выраженными свойствами объектов и контролируемой разреженностью пользовательских историй, сокращенный срез Movies – как мультимедийный случай с отрицательным результатом, сокращенный срез Yelp – как пример сильного популярностного смещения. Семантические признаки строились по текстовым описаниям и метаданным с использованием Sentence-BERT [20]; применялась модель all-MiniLM-L6-v2, доступная по адресу <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>.

В табл. 1 приведены характеристики основных рабочих экспериментальных срезов. Исходный Amazon дополнительно используется как проверочный случай семантического доминирования, поэтому его результаты представлены отдельно в разделе результатов и не смешиваются со сводной характеристикой основных рабочих срезов.

Таблица 1

Характеристики экспериментальных срезов данных

Набор данных	Пользователи	Объекты	Взаимодействия	Среднее на пользователя	Медиана на пользователя
Amazon, минимум 5	1 433	6 841	13 670	9,539	7
Movies, сокращенный срез	265 879	6 387	3 307 732	12,441	9
Yelp, сокращенный срез	895 930	44 733	1 143 545	1,276	1

Примечание: составлена авторами на основе подготовленных экспериментальных срезов данных.

В экспериментальную схему включены BPR-MF [21], LightGCN, NCL, AsymLightGCN, семантический контроль и популярностный контроль. Эксперименты проводились в среде RecBole [22, 23]. Для Amazon, минимум 5, основные результаты агрегировались по шести начальным состояниям генератора случайных чисел: 42, 43, 44, 45, 46 и 47. Для сокращенных срезов Movies и Yelp использовались три начальных состояния: 42, 43 и 44.

AsymLightGCN в этой статье не рассматривается как самостоятельный предмет архитектурного анализа; модель используется как проверяемый графово-семантический метод. LightGCN-часть формирует пользовательский вектор e_u и коллаборативный вектор объекта g_i , семантический вектор

s_i проецируется в латентное пространство $z_i = P_s(s_i)$, а итоговое объектное представление задается смесью

$$\tilde{e}_i = (1 - w_i)g_i + w_i z_i,$$

где вес w_i экспоненциально убывает с ростом степени объекта c_i . При ранжировании используется

$$\hat{y}_{ui} = e_u^\top \tilde{e}_i / (\|\tilde{e}_i\|^\gamma + \varepsilon), \quad \gamma = 0,25.$$

В отличие от LightGCN, модель изменяет объектное представление с учетом семантики и степени объекта, добавляет прототипную семантическую составляющую функции потерь и калибрует норму объектного вектора при ранжировании. Общая функция потерь имеет вид

$$L = e^{-q_{bpr}} (L_{BPR} + \lambda \|\Theta\|_2^2) + q_{bpr} + e^{-q_{sem}} L_{sem} + q_{sem}, \quad (1)$$

где q_{bpr} и q_{sem} – обучаемые коэффициенты балансировки, а L_{sem} – кросс-энтропийная потеря классификации объекта по семантическим прототипам с зависящей от степени температурой и порогом отключения для уже близких прототипу объектов. Точное выражение для w_i и L_{sem} , псевдокод, конфигурации зарегистрированных запусков, программы подготовки данных и векторных представлений, контрольные суммы подготовленных таблиц и статистическая обработка приведены в публичном репозитории воспроизводимости: <https://github.com/nIBOP/asym-lightgcn-reproducibility>. Исходные наборы данных не включаются в пакет реализации; они восстанавливаются по открытым страницам источников с сохранением указанных фильтров и контрольных статистик срезов.

В формуле (1) L_{BPR} обозначает попарную ранжирующую потерю BPR, Θ – совокупность обучаемых параметров модели, а λ – коэффициент L2-регуляризации. Множители $\exp(-q_{bpr})$ и $\exp(-q_{sem})$ задают обучаемые относительные веса двух составляющих; слагаемые q_{bpr} и q_{sem} ограничивают их неограниченное изменение. Такая параметризация соответствует принципу взвешивания задач через обучаемую неопределенность [24].

Для всех обучаемых моделей применялись случайное разбиение 80/10/10, полное ранжирование с исключением обучающих объектов, метрики Recall@10, NDCG@10, Recall@50 и NDCG@50, алгоритм Adam с шагом обучения 0,001, один отрицательный объект на положительный, размерность

векторных представлений 64, максимум 90 эпох, размер пакета 8192, оценка каждые 15 эпох и ранняя остановка после двух проверок без улучшения. Для LightGCN и AsymLightGCN использовались два графовых слоя и $reg_weight=10^{-5}$. Специальные параметры NCL и AsymLightGCN зафиксированы в конфигурациях зарегистрированных запусков в публичном репозитории. Диагностический протокол включает проверку разбиения, расчет основных метрик, семантический и популярностный контроль, стратификацию объектов по популярности и проверку устойчивости ключевых сравнений.

Семантический контроль строит профиль пользователя как среднее семантических векторов объектов из его обучающей истории и ранжирует кандидатов по косинусной близости:

$$p_u = \frac{1}{|I_u^{train}|} \sum_{i \in I_u^{train}} s_i, \quad (2)$$

$$score_{sem}(u, j) = \frac{p_u^\top s_j}{\|p_u\| \|s_j\|},$$

где p_u обозначает семантический профиль пользователя u , I_u^{train} – множество объектов из обучающей истории этого пользователя, $|I_u^{train}|$ – число таких объектов, а s_i и s_j – семантические векторы объектов, построенные по их содержательным признакам. Все известные пользователю объекты исключаются из списка кандидатов до расчета метрик.

Первое выражение в формуле (2) задает профиль как среднее векторных представлений объектов из обучающей истории. Второе выражение вычисляет косинусную близость профиля и объекта: в числителе находится их скалярное произведение, в знаменателе – произведение евклидовых норм. Чем ближе значение к единице, тем выше семантическое соответствие объекта истории пользователя. Этот показатель используется только как контроль и не обучается совместно с проверяемыми моделями.

Популярностный контроль упорядочивает кандидатов по числу обучающих взаимодействий объекта. В обоих контролях используется то же разбиение и то же множество кандидатов, что и для обучаемых моделей. Режим определяется по четырем метрикам Recall@10, NDCG@10, Recall@50 и NDCG@50. Семантическое доминирование фиксировалось, если Semantic-only был не ниже проверяемой модели не менее чем по трем из четырех метрик; популярностное доминирование определялось по тому же правилу для MostPopIndependent. Различие считалось практически малым, если

$$|M_a - M_b| / \max(|M_b|, 10^{-12}) \leq 0,05$$

и интервал парной разности включал ноль; порог 5 % применялся ко всем сравнениям как единая операционная граница практической эквивалентности. Эмпирическое превосходство относительно контролей допускалось при положительной разности не менее чем по трем метрикам, включая Recall@10, причем нижняя граница интервала для Recall@10 должна была быть больше нуля. Если эти условия не выполнялись одинаково для разных метрик, значений K или сегментов объектов, результат относился к смешанному режиму. Отрицательный перенос фиксировался, если новая модель уступала графовому базовому методу не менее чем по трем из четырех метрик.

Результаты исследования и их обсуждение

В табл. 2 показано, как меняется интерпретация при фильтрации Amazon по минимальному числу взаимодействий пользователя. В исходном Amazon семантический контроль слишком силен, поэтому этот вариант не подходит как основной аргумент в пользу графового обучения. Amazon, минимум 5, дает более содержательный сценарий: AsymLightGCN превосходит LightGCN и NCL по метрикам полноты, но семантический контроль остается сильным по раннему ранжированию.

Для проверки устойчивости использовались парные сравнения по начальным состояниям генератора случайных чисел. Для Amazon, минимум 5, дополнительно пересчитаны пользовательские разности метрик: разности усреднялись по шести повторным запускам, после чего по 10 000 бутстрэп-выборок пользователей строился 95 %-й процентильный интервал и выполнялись критерии Уилкоксона и знаков. Для Movies и Yelp приведены 95 %-е t -интервалы среднего парного различия и точный двусторонний знаковый критерий по трем повторным запускам. При трех парах минимальное возможное двустороннее значение этого критерия равно 0,250, поэтому результаты для Movies и Yelp используются как диагностические, а не как самостоятельное строгое доказательство.

В табл. 3 приведена сводная статистическая проверка ключевых сравнений.

Фильтрация Amazon по минимальному числу взаимодействий меняет диагностический режим: в исходном Amazon семантическая близость объектов сама по себе объясняет слишком большую часть качества, а в варианте Amazon, минимум 5, графовые базовые модели становятся сильнее. Поэтому здесь не следует говорить о полном превосходстве модели; корректнее фиксировать область, где семантическая поддержка увеличивает полноту рекомендаций относительно LightGCN и NCL в автономной оценке.

Сокращенный срез Yelp демонстрирует популярностное доминирование: AsymLightGCN выше LightGCN по Recall@10 (0,0109 против 0,0082), но ниже MostPopIndependent (0,0210). Поэтому результаты на данном срезе не подтверждают практическое превосходство графово-семантической модели; здесь прежде всего проявляется популярностное смещение.

Сокращенный срез Movies рассматривается как мультимедийный случай с отрицательным результатом: AsymLightGCN уступает LightGCN по Recall@10 (0,1704 против 0,2024), а семантический контроль практически не работает (0,0015). Этот результат фиксирует границу применимости: наличие содержательных признаков не гарантирует полезности семантического вмешательства, если пользовательские предпочтения лучше восстанавливаются по структуре взаимодействий.

Стратификация уточняет, какие именно объекты попадают в рекомендации. На Amazon, минимум 5, AsymLightGCN получает Recall@10 по редким объектам 0,0409 против 0,0008 у LightGCN и 0,0011 у NCL.

Таблица 2

Результаты на вариантах Amazon

Вариант данных	Метод	Начальные состояния	Recall@10	NDCG@10	Recall@50	NDCG@50
Исходный Amazon	Semantic-only	42	0,2490	0,2201	0,2808	0,2274
Исходный Amazon	LightGCN	42	0,0116	0,0058	0,0423	0,0123
Исходный Amazon	NCL	42	0,0151	0,0072	0,0395	0,0124
Исходный Amazon	AsymLightGCN	42	0,1141	0,0629	0,2164	0,0853
Amazon, минимум 5	MostPop Independent	42/43/44/45/46/47	0,0282 ± 0,0037	0,0133 ± 0,0020	0,0851 ± 0,0066	0,0256 ± 0,0024
Amazon, минимум 5	Semantic-only	42/43/44/45/46/47	0,0441 ± 0,0062	0,0287 ± 0,0053	0,0854 ± 0,0058	0,0378 ± 0,0051
Amazon, минимум 5	BPR-MF	42/43/44/45/46/47	0,0181 ± 0,0042	0,0097 ± 0,0024	0,0415 ± 0,0064	0,0149 ± 0,0027
Amazon, минимум 5	LightGCN	42/43/44/45/46/47	0,0311 ± 0,0036	0,0175 ± 0,0025	0,0878 ± 0,0026	0,0300 ± 0,0019
Amazon, минимум 5	NCL	42/43/44/45/46/47	0,0326 ± 0,0052	0,0177 ± 0,0028	0,0906 ± 0,0041	0,0304 ± 0,0022
Amazon, минимум 5	AsymLightGCN	42/43/44/45/46/47	0,0465 ± 0,0059	0,0232 ± 0,0035	0,1297 ± 0,0070	0,0413 ± 0,0038

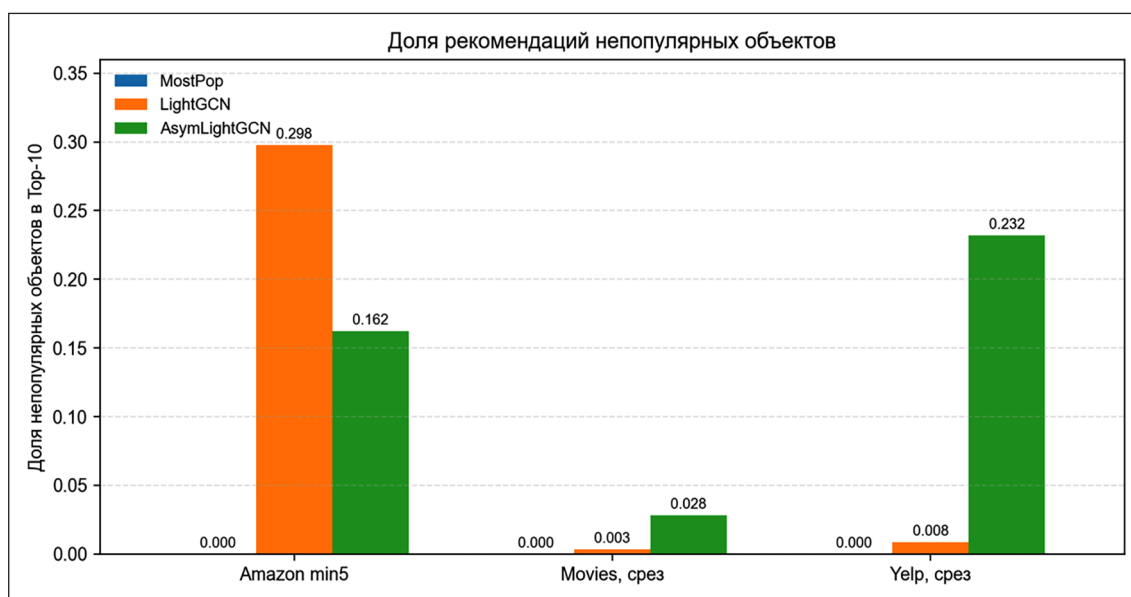
Примечание: составлена авторами по результатам экспериментальной оценки моделей; для шести повторных запусков приведены среднее значение и выборочное стандартное отклонение.

Таблица 3

Сводная статистическая проверка ключевых сравнений

Сценарий	Сравнение	Проверка по метрикам Recall@10; NDCG@10; Recall@50; NDCG@50	Диагностический вывод
Amazon, минимум 5	AsymLightGCN – LightGCN	+0, 0154 [0,0096; 0,0211]; +0, 0057 [0,0022; 0,0093]; +0, 0420 [0,0329; 0,0513]; +0, 0112 [0,0074; 0,0151]	Устойчиво выше графовой базы
Amazon, минимум 5	AsymLightGCN – NCL	+0, 0139 [0, 0080; 0, 0200]; +0, 0055 [0,0021; 0,0090]; +0, 0391 [0, 0300; 0, 0484]; +0, 0109 [0,0073; 0,0146]	Устойчиво выше NCL
Amazon, минимум 5	AsymLightGCN – Semantic-only	+0, 0024 [–0, 0057; 0, 0101]; –0, 0055 [–0, 0111; –0, 0001]; +0, 0443 [0,0328; 0,0559]; +0, 0035 [–0, 0022; 0, 0091]	Смешанный режим: выигрыш по охвату, но не по раннему NDCG
Movies, сокращен- ный срез	AsymLightGCN – LightGCN	–0, 0320 [–0, 0405; –0, 0235]; –0, 0197 [–0, 0250; –0, 0144]; –0, 0443 [–0,0513; –0,0373]; –0, 0227 [–0, 0277; –0, 0178]	Отрицательный перенос во всех ключевых метриках
Yelp, сокращен- ный срез	AsymLightGCN – LightGCN	+0, 0026 [0,0008; 0,0045]; +0, 0013 [–0, 0001; 0, 0028]; +0, 0039 [0, 0030; 0, 0047]; +0, 0017 [0,0008; 0,0026]	Выше LightGCN, но требуется контроль популярности
Yelp, сокращен- ный срез	AsymLightGCN – MostPopIndependent	–0, 0101 [–0, 0114; –0, 0089]; –0, 0050 [–0, 0060; –0, 0040]; –0, 0329 [–0,0347; –0,0311]; –0, 0098 [–0, 0103; –0, 0092]	Популярностное доминирование

Примечание: составлена авторами по результатам статистической обработки. Для Amazon приведены пользовательские бутстрэп-интервалы; для Movies и Yelp – интервалы по трем парным повторным запускам, которые характеризуют только направление различий. Полные таблицы по предметным областям и файлы статистической обработки приведены в публичном репозитории воспроизводимости.



Доля рекомендаций непопулярных объектов в первых десяти рекомендациях
Примечание: составлен авторами по результатам стратифицированного анализа

На сокращенном срезе Movies модель увеличивает долю непопулярных рекомендаций, но не получает релевантных попаданий в этом сегменте. На сокращенном срезе Yelp доля непопулярных рекомендаций у AsymLightGCN выше, чем у LightGCN, однако агрегированное качество остается ниже популярностного контроля.

Доля рекомендаций непопулярных объектов в первых десяти позициях выдачи для рассматриваемых методов представлена на рисунке.

В сводной интерпретации исходный Amazon относится к семантическому доминированию, Amazon, минимум 5, – к смешанному режиму с эмпирическим превосходством по охвату, Movies – к отрицательному переносу, а Yelp – к популярностному доминированию. Поэтому положительный вывод допустим только для Amazon, минимум 5, и только с указанием роли семантического контроля.

Результаты показывают, что один и тот же численный прирост может иметь разные объяснения. На Amazon, минимум 5, AsymLightGCN улучшает графовые базовые модели по полноте ранжирования, но часть раннего ранжирования остается объяснимой семантическим контролем. На Yelp агрегированные метрики связаны с популярностью объектов, а на Movies отрицательный результат фиксирует границу применимости графово-семантического вмешательства. Поэтому протокол работает как предварительный фильтр: он показывает, когда результат можно рассматривать как основание для дальнейшей проверки в пользовательском эксперименте, а когда вывод должен оставаться диагностическим.

Область применимости протокола ограничена несколькими условиями. Во-первых, все эксперименты относятся к автономной оценке и не заменяют проверки пользовательского эффекта в рабочей среде. Во-вторых, применялось случайное разбиение 80/10/10, а не временной протокол. В-третьих, пользовательские бутстрэп-и непараметрические проверки проведены для Amazon, минимум 5; Movies и Yelp проверялись по парным повторным запускам и поэтому интерпретируются осторожнее. В-четвертых, NCL на сокращенных срезах Movies и Yelp не используется как центральное доказательство из-за неполной серии повторных запусков. Наконец, протокол не является причинной оценкой: он диагностирует источники качества, но не восстанавливает контрфактический эффект рекомендационной политики.

Заключение

Предложен диагностический протокол оценки графово-семантических рекомендательных моделей при разреженных данных. Он объединяет полное ранжирование, семантический и популярностный контроль, стратификацию по популярности объектов и проверку устойчивости ключевых сравнений. На исходном Amazon, Amazon с минимальным числом взаимодействий 5, а также на сокращенных срезах Movies и Yelp видно, что агрегированные метрики нельзя интерпретировать без дополнительных контролей. Наиболее содержательный положительный сценарий выявлен на Amazon, минимум 5, где AsymLightGCN улучшает полноту рекомендаций относительно LightGCN и NCL. Результаты на Movies и Yelp, напротив, обозначают границы применимости: в одном случае возникает отрицательный перенос, в другом качество сильнее объясняется популярностным контролем. Работа не утверждает универсальное преимущество конкретной модели, а задает воспроизводимую процедуру различения источников полученного качества.

Список литературы

1. He X., Deng K., Wang X., Li Y., Zhang Y., Wang M. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation // Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020. P. 639–648. DOI: 10.1145/3397271.3401063.
2. Lin Z., Tian C., Hou Y., Zhao W. X. Improving Graph Collaborative Filtering with Neighborhood-enriched Contrastive Learning // Proceedings of the ACM Web Conference 2022. 2022. P. 2320–2329. DOI: 10.1145/3485447.3512104.
3. Anand V., Maurya A. K. A Survey on Recommender Systems Using Graph Neural Network // ACM Transactions on Information Systems. 2025. Vol. 43. Is. 1. Art. 9. P. 1–49. DOI: 10.1145/3694784.
4. Krichene W., Rendle S. On Sampled Metrics for Item Recommendation // Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2020. P. 1748–1757. DOI: 10.1145/3394486.3403226.
5. Sun Z., Yu D., Fang H., Yang J., Qu X., Zhang J., Geng C. Are We Evaluating Rigorously? Benchmarking Recommendation for Reproducible Evaluation and Fair Comparison // Proceedings of the 14th ACM Conference on Recommender Systems. 2020. P. 23–32. DOI: 10.1145/3383313.3412489.
6. Abdollahpouri H., Burke R., Mobasher B. Managing Popularity Bias in Recommender Systems with Personalized Re-Ranking // Proceedings of the Thirty-Second International Florida Artificial Intelligence Research Society Conference (FLAIRS-32) / ed. by R. Barták, K. W. Brawner. AAAI Press, 2019. P. 413–418. URL: <https://cdn.aaai.org/ocs/18199/18199-78818-1-PB.pdf> (дата обращения: 03.06.2026). ISBN 978-1-57735-805-3.
7. Klimashevskaja A., Jannach D., Elahi M., Trattner C. A survey on popularity bias in recommender systems // User Modeling and User-Adapted Interaction. 2024. Vol. 34. P. 1777–1834. DOI: 10.1007/s11257-024-09406-0.
8. Chen J., Dong H., Wang X., Feng F., Wang M., He X. Bias and Debias in Recommender System: A Survey and Future

- Directions // ACM Transactions on Information Systems. 2023. Vol. 41. Is. 3. Art. 67. P. 1–39. DOI: 10.1145/3564284.
9. Melville P., Mooney R. J., Nagarajan R. Content-Boosted Collaborative Filtering for Improved Recommendations // Proceedings of the 18th National Conference on Artificial Intelligence. 2002. P. 187–192. URL: <https://www.cs.utexas.edu/~ml/papers/cbcbf-aaai-02.pdf> (дата обращения: 03.06.2026).
10. Kula M. Metadata Embeddings for User and Item Cold-start Recommendations // Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender Systems (CBRecSys 2015), co-located with the 9th ACM Conference on Recommender Systems / ed. by T. Bogers, M. Koolen. CEUR Workshop Proceedings. 2015. Vol. 1448. P. 14–21. URN тома: urn:nbn:de:0074-1448-5. URL: <https://ceur-ws.org/Vol-1448/paper4.pdf> (дата обращения: 25.08.2026).
11. Yu J., Yin H., Xia X., Chen T., Li J., Huang Z. Self-Supervised Learning for Recommender Systems: A Survey // IEEE Transactions on Knowledge and Data Engineering. 2024. Vol. 36. Is. 1. P. 335–355. DOI: 10.1109/TKDE.2023.3282907.
12. Jing M., Zhu Y., Zang T., Wang K. Contrastive Self-supervised Learning in Recommender Systems: A Survey // ACM Transactions on Information Systems. 2024. Vol. 42. Is. 2. Art. 59. P. 1–39. DOI: 10.1145/3627158.
13. Schnabel T., Swaminathan A., Singh A., Chandak N., Joachims T. Recommendations as Treatments: Debiasing Learning and Evaluation // Proceedings of the 33rd International Conference on Machine Learning. PMLR. 2016. Vol. 48. P. 1670–1679. URL: <https://proceedings.mlr.press/v48/schnabel16.html> (дата обращения: 03.06.2026).
14. Gao C., Zheng Y., Wang W., Feng F., He X., Li Y. Causal Inference in Recommender Systems: A Survey and Future Directions // ACM Transactions on Information Systems. 2024. Vol. 42. Is. 4. Art. 88. P. 1–32. DOI: 10.1145/3639048.
15. Ляликова В. Г., Безрядин М. М. Обзор современных рекомендательных систем // Информационные технологии. 2022. Т. 28. № 9. С. 456–464. URL: <https://journals.econvector.com/1684-6400/article/view/702586> (дата обращения: 03.06.2026). DOI: 10.17587/it.28.456-464.
16. Золотаревский А. Ю., Блохин Н. В. Применение графовых нейронных сетей для построения системы рекомендаций товаров интернет-магазина // Мягкие измерения и вычисления. 2024. № 1–2. С. 22–29. URL: https://s-lib.com/issues/smc_2024_01_2_a3/ (дата обращения: 03.06.2026). DOI: 10.36871/2618-9976.2024.01-2.003.
17. Кулышин Р. С., Сидоров А. А. Обобщенная метрика оценки эффективности алгоритмов рекомендательных систем на основе энтропийного метода // Проблемы управления. 2024. № 4. С. 44–60. URL: <https://journals.rcsi.science/1819-3161/article/view/272865> (дата обращения: 03.06.2026). DOI: 10.25728/pu.2024.4.4.
18. Головинский П. А., Шаталова А. О. Нечеткая рекомендательная система с холодным стартом для выбора траектории обучения // Проблемы управления. 2023. № 6. С. 33–41. URL: <https://journals.rcsi.science/1819-3161/article/view/292103> (дата обращения: 03.06.2026). DOI: 10.25728/pu.2023.6.3.
19. Митрохин М. А., Мартышкин А. И., Ионова Д. Н., Федяшов М. С., Горынина А. В. Метод корректировки релевантности для рекомендательных систем // Современные наукоемкие технологии. 2023. № 12 (ч. 1). С. 60–66. URL: <https://top-technologies.ru/ru/article/view?id=39861> (дата обращения: 03.06.2026). DOI: 10.17513/snt.39861.
20. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics. 2019. P. 3982–3992. URL: <https://aclanthology.org/D19-1410/> (дата обращения: 03.06.2026). DOI: 10.18653/v1/D19-1410.
21. Rendle S., Freudenthaler C., Gantner Z., Schmidt-Thieme L. BPR: Bayesian Personalized Ranking from Implicit Feedback // UAI 2009: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence / ed. by J. A. Bilmes, A. Y. Ng. Montréal, Canada, June 18–21, 2009. AUAI Press, 2009. P. 452–461. [Электронный ресурс]. URL: https://auai.org/uai2009/papers/UAI2009_0139_48141db02b9f0b02bc7158819ebfa2c7.pdf (дата обращения: 03.06.2026). ISBN 978-0-9749039-5-8.
22. Zhao W. X., Mu S., Hou Y., Lin Z., Chen Y., Pan X., Li K., Lu Y., Wang H., Tian C., Min Y., Feng Z., Fan X., Chen X., Wang P., Ji W., Li Y., Wang X., Wen J.-R. RecBole: Towards a Unified, Comprehensive and Efficient Framework for Recommendation Algorithms // Proceedings of the 30th ACM International Conference on Information and Knowledge Management. 2021. P. 4653–4664. DOI: 10.1145/3459637.3482016.
23. Zhao W. X., Hou Y., Pan X., Yang C., Zhang Z., Lin Z., Zhang J., Bian S., Tang J., Sun W., Chen Y., Xu L., Zhang G., Tian Z., Tian C., Mu S., Fan X., Chen X., Wen J.-R. RecBole 2.0: Towards a More Up-to-Date Recommendation Library // Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022. P. 4722–4726. DOI: 10.1145/3511808.3557680.
24. Kendall A., Gal Y., Cipolla R. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018. P. 7482–7491. DOI: 10.1109/CVPR.2018.00781.

Конфликт интересов: Авторы заявляют об отсутствии конфликта интересов.

Conflict of interest: The authors declare that there is no conflict of interest.

Финансирование: Авторы заявляют об отсутствии внешнего финансирования.

Financing: The research was performed without external funding.