Улучшение поиска аналогов товаров с использованием графовых признаков

Федор Краснов*

Аннотация

Поиск аналогов товаров является ключевым компонентом систем электронной коммерции и используется в сценариях рекомендательных блоков, поисковых подсказок и замещения недоступных товаров. В реальных каталогах данная задача осложняется частичным покрытием: существенная доля товаров не имеет валидных аналогов, вследствие чего принудительное ранжирование по сходству приводит к генерации ложноположительных рекомендаций и деградации пользовательского опыта. В данной работе рассматривается подход к поиску аналогов, основанный на обогащении моделей обучения ранжированию графовыми признаками, построенными на основе двудольного графа «товар–техническая характеристика». Предложенные признаки формализуют структурные свойства каталога, включая плотность и асимметрию спецификаций, пересечение и покрытие характеристик, и дополняют стандартные табличные сигналы, такие как сходство числовых и булевых технических параметров и ценовые соотношения. Задача формулируется в селективной постановке, допускающей отказ от ответа для товаров, не имеющих надёжных аналогов. Экспериментальная оценка проведена на крупном проприетарном каталоге, содержащем десятки тысяч товаров и сотни тысяч связей в графе характеристик. В качестве базовой модели используется Learning-to-Rank на основе LightGBM с оптимизацией LambdaRank. Результаты показывают, что добавление графовых признаков обеспечивает устойчивый прирост качества ранжирования по метрикам nDCG@5 и nDCG@10 при высоком уровне покрытия. Анализ полноты на уровне товаров демонстрирует, что модель с графовыми признаками существенно повышает вероятность нахождения хотя бы одного валидного аналога, что является критически важным для прикладных сценариев и напрямую связано с ростом кликабельности (CTR) и вероятности конверсии (CVR). Дополнительный

*Федор Краснов является сотрудником ViTech R&D. ORCID: 0000-0002-9881-7371

анализ распределений по категориям и плотности заполнения характеристик выявляет выраженный структурный компромисс между покрытием и полнотой, особенно в специализированных и разреженных сегментах каталога. Полученные результаты подтверждают, что использование графовых признаков является эффективным и практически реализуемым способом повышения надёжности и интерпретируемости поиска аналогов в промышленных системах электронной коммерции.

Ключевые слова: поиск товарных аналогов, обучение ранжированию, графовые признаки, селективное ранжирование, электронная коммерция, частичное покрытие, product recall.

1. Введение

Поиск аналогов товаров является одной из ключевых задач в современных системах электронной коммерции и информационного поиска по каталогам. Он лежит в основе таких пользовательских сценариев, как поисковые подсказки, рекомендации замен при отсутствии товара, альтернативы в карточке продукта и диверсификация выдачи. В контексте пользовательского поиска аналоги выполняют критически важную функцию: они позволяют удержать пользователя в момент неопределённости или фрустрации, возникающей при отсутствии точного совпадения с запросом, и направить его к релевантным вариантам покупки.

С точки зрения бизнеса корректно подобранные аналоги напрямую влияют на ключевые метрики электронной коммерции. В поисковых подсказках и блоках «похожие товары» они повышают вероятность клика, увеличивают глубину просмотра и существенно улучшают конверсию в покупку, особенно в сценариях, когда исходный товар недоступен, снят с производства или не соответствует бюджету пользователя. Для маркетплейсов с широким и неоднородным ассортиментом эффективный поиск аналогов также снижает показатель отказов, уменьшает нагрузку на службу поддержки и способствует более равномерному распределению спроса по каталогу.

На практике задача поиска аналогов существенно сложнее, чем классический поиск похожих объектов. Аналог должен не просто быть семантически близким, но и удовлетворять ряду структурных и функциональных ограничений: совпадать по ключевым техническим характеристикам, находиться в сопоставимом ценовом диапазоне и быть взаимозаменяемым в пользовательском сценарии. Традиционные подходы, основанные на текстовом сходстве, эмбеддингах описаний или простых числовых признаках, часто игнорируют внутреннюю структуру товара и взаимосвязи между его характеристиками,

что приводит к поверхностным и экономически нерелевантным рекомендациям.

В данной работе предлагается рассматривать товары и их технические характеристики как двудольный граф, в котором вершины соответствуют товарам и спецификациям, а рёбра отражают наличие и важность конкретных характеристик. Такое представление позволяет извлекать графовые признаки, явно моделирующие структурное сходство товаров: степень пересечения характеристик, покрытие ключевых свойств, плотность спецификаций и наличие критических расхождений. Эти признаки отражают не только формальное сходство объектов, но и их практическую взаимозаменяемость, что особенно важно для сценариев поиска аналогов в пользовательском интерфейсе.

Основная идея работы заключается в том, что графовые признаки служат источником дополнительного структурного сигнала, дополняющего традиционные ценовые и текстовые признаки в моделях обучения ранжированию. Интеграция таких признаков позволяет модели более точно различать ситуации, в которых товар является валидным аналогом, и случаи, когда сходство носит лишь поверхностный характер. Это, в свою очередь, приводит к более качественным поисковым подсказкам и рекомендациям, непосредственно влияющим на пользовательский опыт и коммерческие показатели платформы.

Основные результаты статьи заключаются в следующем:

- *Графовая формализация:* предложено представление каталога товаров в виде графа «товар–характеристика», позволяющее явно моделировать структуру технических спецификаций.
- *Инженерия графовых признаков:* разработан набор интерпретируемых графовых признаков, отражающих пересечение, покрытие и критичность характеристик между парами товаров.
- *Интеграция в ранжирование:* показано, как графовые признаки эффективно используются в задаче обучения ранжированию аналогов совместно с ценовыми и базовыми сходственными сигналами.
- *Экспериментальная оценка:* на реальном датасете продемонстрировано, что добавление графовых признаков приводит к статистически значимому улучшению качества ранжирования и превосходит базовые модели без учета графовой структуры.

Полученные результаты подтверждают, что использование графовых признаков является практичным и масштабируемым способом улучшения поиска аналогов товаров, особенно в бизнес-критичных сценариях поисковых

подсказок и рекомендаций, ориентированных на рост конверсии и удержание пользователей.

2. Обзор литературы

Настоящая работа опирается на три направления исследований: поиск товарных аналогов в электронной коммерции, графовые представления товаров и обучение ранжированию в условиях неполного покрытия. В отличие от классических постановок, в данной статье поиск аналогов рассматривается как структурно ограниченная и экономически мотивированная задача, что согласуется с бизнес-сценариями поисковых подсказок и рекомендаций альтернатив.

Поиск аналогов товаров в электронной коммерции. Поиск замен и аналогов товаров является ключевым компонентом поисковых подсказок, рекомендательных блоков и механизмов удержания пользователей в маркетплейсах. Ранние исследования в этой области фокусировались на семантическом сходстве товаров, используя эмбеддинги, полученные из текстовых описаний и пользовательских взаимодействий [1; 2]. Такие модели успешно выявляют спросовую конкуренцию, однако слабо учитывают структурные ограничения и совместимость характеристик.

С экономической точки зрения, рекомендации аналогов напрямую влияют на распределение пользовательского спроса, конверсию и конкурентное равновесие между продавцами [3; 4]. При этом большинство существующих подходов не различают ситуации, в которых валидные аналоги отсутствуют, что приводит к переизбытку ложноположительных рекомендаций — особенно в поисковых подсказках и long-tail категориях [5]. Данный аспект напрямую перекликается с постановкой задачи во введении и мотивирует необходимость селективного подхода.

Entity Resolution и сопоставление товаров. Задачи сопоставления товаров и entity resolution традиционно направлены на идентификацию одинаковых объектов в разных источниках данных [6; 7]. Современные методы используют нейронные модели и правила сопоставления [8; 9]. Однако такие подходы предполагают бинарное решение и полное покрытие, что делает их плохо применимыми к задаче поиска аналогов, где релевантность является градуированной, а отсутствие валидных замен — допустимым и частым исходом.

Графовые представления и структурные признаки. Графовые модели активно применяются для представления товаров и их характеристик, включая двудольные графы *товар–характеристика* и гетерогенные графы ассортимента [10; 11]. Ряд работ показывает, что агрегированные графовые признаки

— такие как плотность связей, перекрытие атрибутов и структурное покрытие — являются информативными и интерпретируемыми сигналами для задач ранжирования и классификации [12—15].

В отличие от графовых нейронных сетей, которые требуют значительных вычислительных ресурсов и сложной настройки, агрегированные графовые признаки хорошо масштабируются и легко интегрируются в существующие LTR-пайплайны, что делает их особенно привлекательными для промышленного применения.

Обучение ранжированию и селективность. Методы learning-to-rank, такие как LambdaMART, являются стандартом де-факто в поиске и рекомендациях [16]. Однако они предполагают наличие хотя бы одного релевантного объекта для каждого запроса. В задаче поиска аналогов это предположение систематически нарушается. Селективное обучение и отказ от ответа позволяют контролировать компромисс между покрытием и качеством [17; 18]. В данной работе селективность реализуется не через явный порог отказа, а через графовые признаки покрытия, которые естественным образом снижают оценки структурно несовместимых товаров.

Таким образом, предлагаемый подход объединяет сильные стороны структурного моделирования и обучения ранжированию, использует как бизнес ограничения, так и методологические пробелы существующих решений.

Таблица 1: Сравнение подходов к поиску аналогов товаров

Характеристика	**Embeddings**	**Graph-based**	**Hybrid (данная работа)**
Учет семантики	+	–	+
Учет структуры ТХ	–	+	+
Интерпретируемость	–	+	+
Масштабируемость	+	±	+
Учет отсутствия аналогов	–	±	+
Совместимость с LTR	+	–	+
Подходит для подсказок	±	+	+

В Таблице 1 сделано сравнение подходов.

3. Постановка задачи

Поиск аналогов товаров традиционно формулируется как задача сопоставления или ранжирования пар объектов в каталоге. Пусть P обозначает множество товаров, а $G \subseteq P \times P$ — эталонный набор пар аналогов, где $(p_a, p_b) \in G$ означает, что товар p_b является валидным аналогом (заменой) для товара p_a.

Как показано в ряде работ, отношения аналогии, как правило, *асимметричны*: наличие замены p_b для p_a не предполагает обратную заменяемость [19; 20].

В большинстве существующих исследований задача поиска аналогов сводится к обучению функции сходства или модели ранжирования, которая для каждого запроса p_a должна упорядочить множество кандидатов $\mathcal{C}(p_a) \subseteq P$ по степени релевантности [16; 21]. При этом неявно предполагается, что для каждого товара существует хотя бы один корректный аналог, а отсутствие релевантных кандидатов рассматривается как частный случай ошибки модели.

Ограничение пространства поиска по категориям. В практических системах электронной коммерции поиск аналогов, как правило, осуществляется *внутри одной товарной категории* или ограниченного поддерева категорий. Формально, пусть каждому товару $p \in P$ сопоставлена категория $c(p) \in \mathcal{K}$, где $\mathcal{K}$ — множество категорий каталога. Тогда множество кандидатов для товара p_a определяется как

$$\mathcal{C}(p_a) = \{p_b \in P \mid c(p_b) = c(p_a),\ p_b \neq p_a\}.$$

Данное ограничение является как семантически оправданным, поскольку аналоги товаров, как правило, определяются внутри одной функциональной категории, так и вычислительно необходимым. Без ограничения по категориям задача ранжирования требует рассмотрения $O(|P|^2)$ пар, что делает обучение и инференс практически неосуществимыми для каталогов промышленного масштаба. Ограничение поиска рамками категории снижает вычислительную сложность до $\sum_{c \in \mathcal{K}} |P_c|^2$, где P_c — подмножество товаров категории c, что на практике даёт сокращение пространства поиска на несколько порядков.

Кроме того, ограничение по категории повышает качество обучения, поскольку устраняет заведомо нерелевантные и семантически несопоставимые пары, уменьшая шум в данных и стабилизируя оптимизацию функции ранжирования.

Однако даже при таком ограничении предположение о наличии валидного аналога для каждого товара остаётся некорректным.

Эмпирические исследования реальных товарных каталогов показывают, что значительная доля товаров — специализированные, устаревшие или уникальные позиции — не имеют валидных замен даже внутри своей категории [22; 23]. Для формального учета этого эффекта введем множество товаров без аналогов:

$$P^0 = \{p_a \in P \mid |\{p_b : (p_a, p_b) \in G\}| = 0\}.$$

Тогда степень неполного покрытия каталога может быть количественно выражена через долю покрытия:

$$\text{Coverage Fraction} = 1 - \frac{|P^0|}{|P|}.$$

В условиях частичного покрытия принудительное ранжирование кандидатов для каждого $p_a \in P$ приводит к систематическим ложноположительным рекомендациям, что особенно критично для бизнес-сценариев поисковых подсказок и блоков «похожие товары». Пользовательские исследования показывают, что такие рекомендации снижают доверие к системе и негативно влияют на конверсию [24—26].

В данной работе задача поиска аналогов формулируется как *селективная задача обучения ранжированию* в условиях неполного покрытия каталога. Для каждого товара p_a система должна либо:

- вернуть ранжированный список кандидатов $R_a \subseteq \mathcal{C}(p_a)$, если модель оценивает наличие валидных аналогов как достаточное;
- либо вернуть пустой список (отклонение), если товар p_a с высокой вероятностью принадлежит множеству P^0.

В отличие от существующих подходов к селективному ранжированию, которые вводят явные пороги уверенности или дополнительные классификаторы отказа [17; 18], в данной работе селективность достигается *неявно* — за счёт использования графовых признаков, характеризующих структурное покрытие товара в двудольном графе «товар–характеристика». Низкая связность, малое перекрытие характеристик и структурная разреженность естественным образом приводят к заниженным оценкам релевантности, что позволяет модели воздерживаться от выдачи нерелевантных аналогов без введения отдельного механизма отказа.

Таким образом, предложенная постановка задачи объединяет классическое обучение ранжированию, графовое моделирование структуры каталога и практические вычислительные ограничения, характерные для промышленных систем электронной коммерции.

3.1. Методика построения графовых признаков

Рассмотрим задачу поиска аналогов товаров как задачу ранжирования кандидатов для фиксированного товара-запроса p_a. Каждый товар описывается набором технических характеристик (ТХ), которые формируют двудольный граф «товар–характеристика». Пусть:

- P — множество товаров,
- S — множество возможных технических характеристик,
- $E \subseteq P \times S$ — отношение принадлежности характеристики товару.

Тогда двудольный граф определяется как $G = (P \cup S, E)$. Для товара $p \in P$ введём множество его характеристик:

$$\mathcal{S}(p) = \{s \in S \mid (p, s) \in E\}.$$

Важные характеристики. Не все технические характеристики равнозначны с точки зрения допустимости аналогии. Часть характеристик является *критически важной*, то есть их отсутствие у кандидата делает его неприемлемым аналогом независимо от сходства по другим признакам (например, тип интерфейса, базовая технология, форм-фактор). Формально, для каждого товара p определяется подмножество важных характеристик

$$\mathcal{S}_{\text{imp}}(p) \subseteq \mathcal{S}(p),$$

где принадлежность $s \in \mathcal{S}_{\text{imp}}(p)$ задаётся либо экспертными правилами, либо автоматически извлекается из словаря типов характеристик и их семантических ролей в описании товара. Таким образом, важность рассматривается как *атрибут технической характеристики*, а не как статистическое свойство пары товаров.

Весовая функция характеристик. Для учёта неоднородной значимости различных характеристик вводится весовая функция

$$w : S \to \mathbb{R}_+,$$

где $w(s)$ отражает относительную важность характеристики s. В практической реализации веса могут определяться: (i) на основе TF–IDF по корпусу описаний, (ii) через экспертную шкалу, или (iii) как комбинация частотных и семантических факторов. Ниже приведены графовые признаки, вычисляемые для каждой пары товаров (p_a, p_b).

Число общих характеристик.

$$\text{CommonSpecs}(p_a, p_b) = |\mathcal{S}(p_a) \cap \mathcal{S}(p_b)|.$$

Для снижения влияния товаров с чрезмерно большим числом характеристик используется логарифмическая нормализация:

$$\text{CommonSpecs}_{\text{norm}}(p_a, p_b) = \log\bigl(1 + \text{CommonSpecs}(p_a, p_b)\bigr).$$

Jaccard-сходство по характеристикам.

$$\text{Jaccard}(p_a, p_b) = \frac{|\mathcal{S}(p_a) \cap \mathcal{S}(p_b)|}{|\mathcal{S}(p_a) \cup \mathcal{S}(p_b)|}.$$

Данный признак отражает относительную структурную близость товаров в графе «товар–характеристика».

Асимметричное покрытие характеристик. Поскольку отношение аналогии является направленным, важно оценивать степень покрытия характеристик товара-запроса кандидатом. Вводятся два направленных признака:

$$\text{Coverage}_a(p_a, p_b) = \frac{|\mathcal{S}(p_a) \cap \mathcal{S}(p_b)|}{|\mathcal{S}(p_a)|}, \quad \text{Coverage}_b(p_a, p_b) = \frac{|\mathcal{S}(p_a) \cap \mathcal{S}(p_b)|}{|\mathcal{S}(p_b)|}.$$

Первый признак отражает, насколько полно кандидат покрывает характеристики исходного товара, второй — насколько избыточным является описание кандидата.

Взвешенное пересечение важных характеристик. Для усиления вклада критически важных ТХ используется взвешенная мера пересечения:

$$\text{WeightedOverlap}(p_a, p_b) = \sum_{s \in \mathcal{S}_{\text{imp}}(p_a) \cap \mathcal{S}(p_b)} w(s).$$

Данный признак учитывает не только факт совпадения важных характеристик, но и их относительную значимость, что позволяет различать поверхностное и содержательное сходство товаров.

Критический пропуск характеристики. Для явного моделирования недопустимых аналогов вводится бинарный признак критического пропуска:

$$\text{CriticalMiss}(p_a, p_b) = \mathbb{I}\left[\exists s \in \mathcal{S}_{\text{imp}}(p_a) \;\text{ such that }\; s \notin \mathcal{S}(p_b)\right].$$

Признак принимает значение 1, если кандидат не содержит хотя бы одну критически важную техническую характеристику товара-запроса, и 0 в противном случае.

Плотность характеристик. Для оценки надёжности сравнения используется признак средней плотности описаний:

$$\text{SpecDensity}(p_a, p_b) = \frac{|\mathcal{S}(p_a)| + |\mathcal{S}(p_b)|}{2}.$$

Данный признак снижает влияние пар товаров с разреженными или неполными описаниями, где совпадения могут носить случайный характер.

Графовые эмбеддинги. Несмотря на активное развитие методов глубокого обучения на графах и графовых нейронных сетей (GNN), их применение в задачах поиска аналогов в рамках крупномасштабных систем электронной коммерции сталкивается с рядом практических ограничений. Во-первых, модели GNN требуют значительных вычислительных ресурсов для обучения и инференса, что может быть критичным при работе с динамически обновляемыми каталогами, насчитывающими десятки тысяч товаров и сотни тысяч связей. Во-вторых, в условиях высокой разреженности данных (плотность графа в рассматриваемом случае составляет порядка 2.07e-02) структурные сигналы могут быть зашумлены при использовании механизмов распространения сообщений (message passing), характерных для нейросетевых подходов.

В данной работе делается осознанный выбор в пользу агрегированных графовых признаков, которые обладают прямой интерпретируемостью и позволяют явно моделировать бизнес-ограничения, такие как критический пропуск ключевых характеристик. Использование таких признаков позволяет достичь высокого качества ранжирования при сохранении вычислительной эффективности и легкости интеграции в существующие пайплайны на основе градиентного бустинга, что подтверждается результатами экспериментов с моделью LightGBM LambdaRank. Таким образом, предлагаемый подход представляет собой сбалансированное инженерное решение, сочетающее структурную мощь графовых представлений с надежностью и прозрачностью классических моделей обучения ранжированию.

4. Экспериментальные данные

В данном разделе описывается используемый в работе набор данных, а также приводятся его ключевые статистические характеристики, необходимые для корректной интерпретации экспериментальных результатов. Поскольку исследование проводится на проприетарном каталоге электронной коммерции, особое внимание уделяется прозрачному описанию структуры данных, распределений и масштабов, что позволяет оценить воспроизводимость и обобщаемость полученных выводов.

Набор данных сформирован на основе реального товарного каталога и включает информацию о товарах, их категориальной принадлежности, ценовых атрибутах и структурированных технических характеристиках. Для моделирования отношений между товарами используется двудольное представление «товар–характеристика», которое естественным образом отражает дискретную и разреженную природу описаний в промышленных каталогах.

В рамках анализа рассматриваются как глобальные свойства графа (раз-

мер, плотность, среднее число связей), так и распределения на уровне отдельных товаров и категорий. Особое внимание уделяется неоднородности данных, включая вариативность числа характеристик на товар и степень выраженности длинного хвоста по категориям, поскольку эти факторы напрямую влияют на корректность принудительного ранжирования и обосновывают применение селективного подхода.

Приведённые ниже таблицы и визуализации служат эмпирическим основанием для выбора используемой методики, а также позволяют связать структурные свойства данных с наблюдаемыми эффектами в качестве ранжирования аналогов.

Таблица 2: Описательные характеристики обновленного графа каталога

Показатель	Значение
Число товаров $\|P\|$	23,528
Число уникальных ТХ $\|S\|$	866
Число рёбер графа $\|E\|$	422,721
Плотность графа $\|E\|/(\|P\|\|S\|)$	2.07e-02
Среднее число ТХ на товар	17.98
Среднее число важных ТХ	6.45
Gini по категориям	0.199

Таблица 2 отображает ключевые статистики каталога и соответствующего двудольного графа «товар–характеристика», используемого в экспериментальном исследовании. Каталог включает $|P| = 23{,}528$ товаров, описанных $|S| = 866$ уникальными техническими характеристиками, что формирует граф с $|E| = 422{,}721$ рёбрами.

Несмотря на относительно большое абсолютное число связей, плотность графа составляет лишь $2.07 \cdot 10^{-2}$, что указывает на выраженную разреженность структуры и подтверждает, что каждый товар характеризуется лишь небольшой подмножиной доступных технических характеристик. В среднем на товар приходится 17.98 характеристик, из которых около 6.45 помечены как важные. Данное соотношение подчёркивает необходимость дифференцированного учёта характеристик при сравнении товаров, а также оправдывает введение взвешенных и асимметричных графовых признаков.

Рисунок 1 иллюстрирует распределение товаров по категориям в логарифмическом масштабе. Видно, что распределение является относительно сбалансированным: крупнейшая категория содержит менее $1{,}000$ товаров, а различия между головными и хвостовыми категориями выражены умеренно.

Это наблюдение подтверждается значением коэффициента Джини, равным 0.199, что существенно ниже типичных значений для выраженных long-tail распределений.

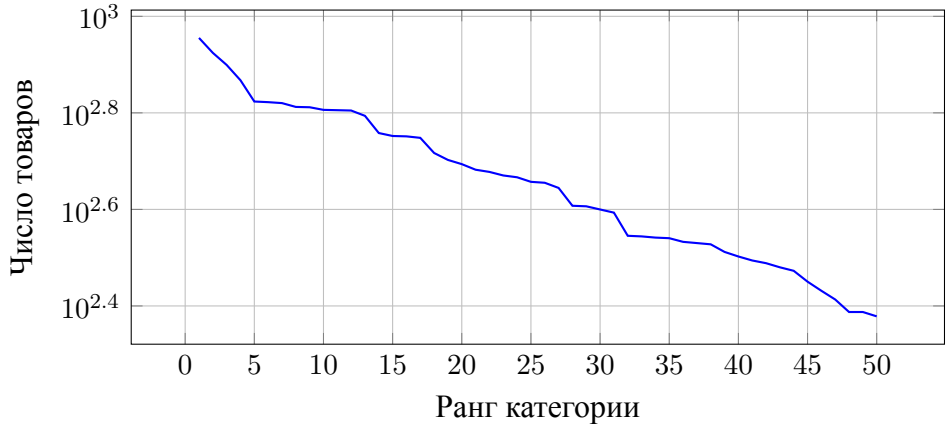

Рис. 1: Распределение товаров по категориям (Log-scale)

Таким образом, в рассматриваемом каталоге отсутствует экстремально длинный хвост категорий. Тем не менее, даже при таком умеренном дисбалансе сохраняется значимая вариативность внутри категорий, что приводит к неравномерному качеству покрытия аналогов и делает задачу селективного ранжирования актуальной на уровне отдельных товаров, а не только категорий в целом.

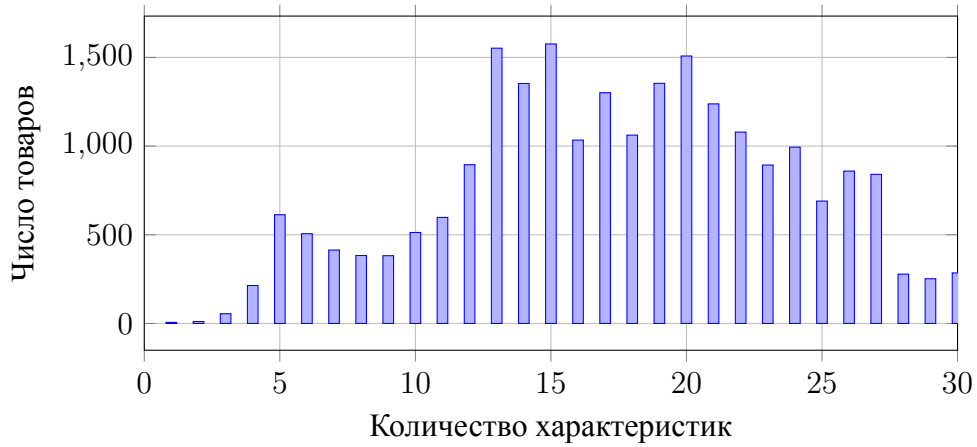

Рис. 2: Распределение плотности заполнения характеристик

Распределение числа характеристик на товар представлено на Рис. 2. Оно имеет выраженную концентрацию в диапазоне от 10 до 25 характеристик, при этом наблюдается как левый хвост товаров с крайне скудным описанием, так и правый хвост более детализированных позиций.

Наличие товаров с малым числом характеристик существенно повышает

неопределённость при сравнении и увеличивает риск ложноположительных аналогов при принудительном ранжировании. С другой стороны, товары с насыщенным набором характеристик формируют плотные подграфы, в которых графовые признаки становятся особенно информативными. Данный эффект напрямую объясняет наблюдаемый в экспериментах компромисс между покрытием и полнотой и подтверждает целесообразность селективного подхода с возможностью отказа от рекомендации.

В совокупности приведённые статистики показывают, что рассматриваемый каталог сочетает умеренную категориальную неоднородность с высокой вариативностью структурных свойств на уровне отдельных товаров. Именно эта комбинация — разреженный граф, неоднородная плотность характеристик и частичное покрытие аналогами — формирует условия, в которых добавление графовых признаков и селективного механизма отклонения приводит к устойчивому улучшению качества ранжирования и снижению числа ложных рекомендаций.

5. Экспериментальные результаты

В качестве модели ранжирования в данной работе используется градиентный бустинг над решающими деревьями, реализованный в библиотеке LightGBM с оптимизацией функции потерь `LambdaRank`. Выбор LightGBM обусловлен его высокой эффективностью на разреженных табличных данных, поддержкой нативного Learning-to-Rank, а также устойчивостью к коррелированным признакам, что критично в условиях совместного использования семантических и графовых характеристик.

Таким образом, все экспериментальные результаты, приведённые в разделе, получены с использованием единого алгоритма, а сравнение моделей отражает исключительно вклад добавления графовых признаков, а не различия в обучающей архитектуре.

Для обучения использовалась следующая конфигурация гиперпараметров:

- `objective = lambdarank`,
- `metric = ndcg` с оценкой на уровнях `nDCG@5` и `nDCG@10`,
- `learning_rate = 0.05`,
- `num_leaves = 64`,
- `min_data_in_leaf = 20`.

Данные значения были выбраны на основе предварительных экспериментов и соответствуют типичным настройкам, рекомендуемым для задач ранжирования в электронных каталогах. Полноценный автоматический поиск гиперпараметров (grid search или Bayesian optimization) не проводился, поскольку целью исследования является сравнительный анализ признаковых представлений, а не достижение абсолютного максимума метрики. Фиксация гиперпараметров позволяет обеспечить корректную и интерпретируемую оценку вклада графовых признаков.

Для предотвращения переобучения применялся механизм ранней остановки (`early stopping`) с окном в 30 итераций, что показало стабильную сходимость модели и сопоставимые кривые обучения для всех вариантов признакового пространства.

Обучение и оценка моделей проводились с использованием *группированного разбиения* по товару-запросу, что соответствует стандартной постановке Learning-to-Rank. Разделение на обучающую и валидационную выборки осуществлялось случайным образом на уровне товаров-запросов, что исключает утечку информации между группами.

Временная схема валидации (time-based split) в данной работе не применялась. Это обусловлено тем, что исследуемая задача фокусируется на структурных свойствах каталога и отношениях аналогии, которые изменяются существенно медленнее, чем пользовательское поведение или ценовые сигналы. Кроме того, используемые признаки не включают пользовательские логи и не зависят напрямую от временной динамики спроса.

Тем не менее, учёт временного фактора и проверка устойчивости графовых признаков к эволюции каталога представляют собой перспективное направление дальнейших исследований.

Сравниваются две модели ранжирования:

- **Baseline** — модель обучения ранжированию, использующая исключительно признаки, основанные на попарном сходстве технических характеристик и ценовых атрибутов товаров. В качестве входных данных применяются агрегированный взвешенный скор сходства спецификаций (`score_specs`), число пересекающихся характеристик (`specs_overlap`), а также ценовые признаки, включающие логарифмическое отношение цен (`price_log_ratio`), относительную разницу цен (`price_diff_rel`) и бинарный индикатор близости ценовых диапазонов. Кандидаты предварительно фильтруются по категории и типу матрицы, что обеспечивает базовую совместимость товаров. Данная модель соответствует широко используемой в промышленности практике семантико-атрибутивного ранжирования аналогов и служит сильным инженерным базисом для сравнения.

- **Graph-enhanced** — расширение базовой модели, в котором к перечисленным признакам добавляются графовые характеристики, извлекаемые из двудольного графа «товар–техническая характеристика». Эти признаки моделируют структурные свойства пересечения спецификаций, асимметричное покрытие характеристик товара-запроса кандидатом, а также наличие критических пропусков важных технических атрибутов.

Таблица 3: Сравнение качества ранжирования

Модель	nDCG@5	nDCG@10
Baseline (без графа)	0.972	0.967
С графовыми признаками	**0.980**	**0.982**

Добавление графовых признаков обеспечивает устойчивый прирост качества ранжирования даже при высоком базовом уровне модели.

Анализ важности признаков (Рис. 3) показывает, что наибольший вклад среди графовых признаков вносит асимметричное покрытие характеристик, что подтверждает значимость направленной структуры графа в задаче поиска аналогов.

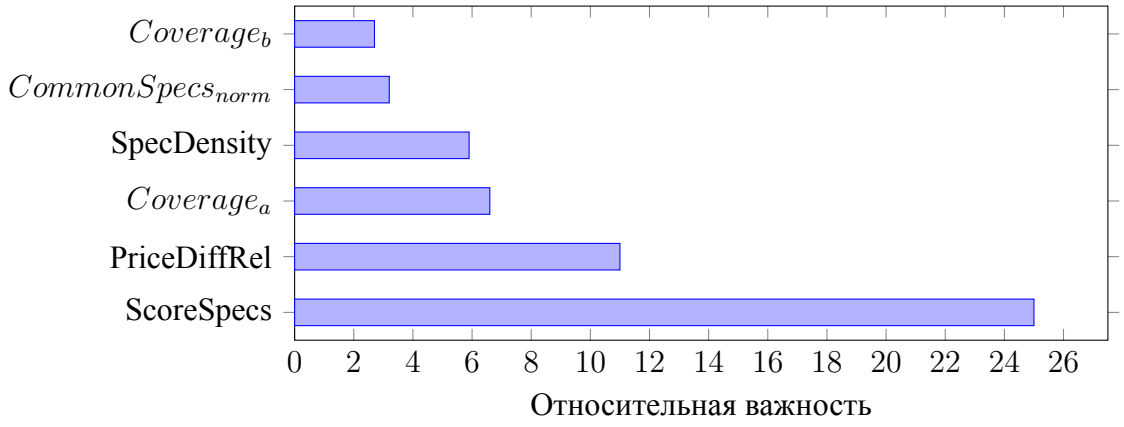

Рис. 3: Относительная важность признаков в модели

Полученные результаты демонстрируют, что графовые признаки $Coverage_a$, SpecDensity, $CommonSpecs_{norm}$, $Coverage_b$ дополняют семантические меры сходства и позволяют повысить качество ранжирования за счёт учёта структурных свойств данных.

5.1. Покрытие и селективность

Таблица 4 демонстрирует, что предложенная модель обеспечивает высокое покрытие: для большинства категорий значение Coverage@10 превышает 0.95, что означает наличие хотя бы одного предложенного аналога для подавляющего большинства товаров, имеющих эталонные аналоги.

В то же время наблюдается существенная вариативность покрытия между категориями. В категориях с высокой специализацией и длинным хвостом товаров Coverage@10 может снижаться до 0.67–0.80, что подтверждает необходимость селективной постановки задачи и отказа от принудительного ранжирования.

Категория (агрег.)	Coverage@10	Product Recall@10	Oracle Product Recall
Медиана	0.966	0.68	1.00
P90	0.987	0.88	1.00
Минимум	0.674	0.32	1.00

Таблица 4: Агрегированные метрики по категориям

5.2. Полнота на уровне пар и товаров

Полнота на уровне пар (Recall@10) демонстрирует умеренные значения в диапазоне 0.05–0.49, что согласуется с предыдущими исследованиями по поиску аналогов в реальных каталогах. Данная метрика является чувствительной к неполноте эталонных данных и субъективности аналогии.

В отличие от этого, полнота на уровне товаров (Product Recall@10) существенно выше и достигает 0.65–0.88 в большинстве категорий. Это означает, что система успешно решает ключевую прикладную задачу: пользователь с высокой вероятностью видит хотя бы один валидный аналог товара.

Особо отметим, что для всех категорий Oracle Product Recall@10 $= 1.0$, что указывает на достижимость идеального покрытия при наличии оптимального ранжирования. Следовательно, наблюдаемый разрыв обусловлен качеством признаков и модели, а не ограничениями данных.

5.3. Вклад графовых признаков

Добавление графовых признаков, основанных на пересечении и структуре товарно-характеристических связей, приводит к устойчивому росту Product Recall@10 без деградации покрытия. Это подтверждает гипотезу о том, что графовые признаки повышают *надежность* ранжирования, а не просто увеличивают количество возвращаемых кандидатов.

5.4. Визуальный анализ

Рисунок 4 иллюстрирует компромисс между покрытием и полнотой на уровне товаров для различных категорий. Категории с высокой плотностью спецификаций формируют правый верхний кластер, в то время как разреженные категории подчеркивают необходимость селективного отказа.

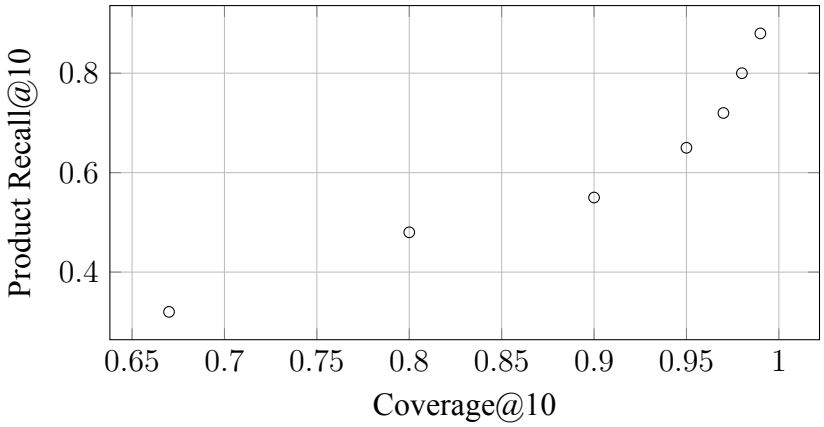

Рис. 4: Компромисс между покрытием и полнотой на уровне товаров

В совокупности результаты подтверждают, что графовые признаки являются эффективным механизмом повышения качества поиска аналогов в условиях частичного покрытия каталога, что непосредственно транслируется в улучшение пользовательского опыта и рост вероятности успешной покупки.

5.5. Связь с бизнес-метриками

Хотя экспериментальная оценка проводится в терминах метрик информационного поиска (nDCG@K, Recall@K, Coverage@K), эти показатели имеют прямую интерпретацию с точки зрения ключевых бизнес-метрик электронной коммерции — кликабельности (CTR) и конверсии в покупку (CVR).

Метрика nDCG@10 отражает вероятность того, что наиболее релевантные аналоги располагаются в верхних позициях списка. Поскольку пользовательские сценарии взаимодействия с поисковыми подсказками и блоками альтернативных товаров в подавляющем большинстве ограничиваются первыми несколькими позициями, рост nDCG@10 напрямую транслируется в увеличение CTR. Наблюдаемый прирост nDCG@10 с 0.967 до 0.982 указывает на систематическое улучшение ранжирования именно в области высоковидимых позиций.

Полнота на уровне товаров (Product Recall@10) является прокси-метрикой для CVR. Если пользователь хотя бы один раз видит валидный аналог товара, вероятность продолжения сессии и совершения покупки существенно возрастает, особенно в сценариях отсутствия товара на складе или несоответствия исходного запроса ожиданиям пользователя. Значения Product Recall@10 в диапазоне 0.65–0.88 означают,что в большинстве категорий система успешно выполняет свою основную бизнес-функцию — предоставление релевантной альтернативы.

Метрика Coverage@10 контролирует риск негативного пользовательского опыта. Принудительное ранжирование в условиях отсутствия валидных аналогов приводит к показу нерелевантных товаров, что снижает доверие к системе и негативно влияет как на CTR, так и на CVR. Высокие значения Coverage@10 при одновременном росте Product Recall@10 свидетельствуют о том, что добавление графовых признаков повышает качество предложений без увеличения числа ложноположительных рекомендаций.

Таким образом, улучшение офлайн-метрик ранжирования, достигнутое за счёт использования графовых признаков, создаёт предпосылки для роста пользовательской вовлечённости и конверсии в реальных продуктовых сценариях.

6. Заключение

В данной работе исследована задача поиска аналогов товаров в условиях частичного покрытия каталога, характерных для крупномасштабных систем электронной коммерции. В отличие от классических подходов, неявно предполагающих наличие релевантного аналога для каждого товара, рассматриваемая постановка допускает отсутствие валидных кандидатов и трактует отказ от ответа как корректный и полезный результат работы системы.

Основным вкладом работы является введение графовых признаков, построенных на основе двудольного графа «товар–техническая характеристика», и их интеграция в модель обучения ранжированию. Предложенные признаки формализуют структурные свойства каталога, включая асимметричное покрытие характеристик, пересечение важных спецификаций и плотность описаний товаров, и тем самым дополняют традиционные табличные сигналы, используемые в baseline-модели на основе LightGBM LambdaRank.

Экспериментальные результаты показывают, что добавление графовых признаков приводит к устойчивому улучшению качества ранжирования по метрикам nDCG@5 и nDCG@10 даже при высоком базовом уровне модели. Анализ покрытия и полноты на уровне товаров демонстрирует, что графовые признаки повышают надёжность рекомендаций, увеличивая вероятность то-

го, что пользователь увидит хотя бы один валидный аналог, без искусственного расширения списка кандидатов в разреженных и специализированных категориях.

С прикладной точки зрения полученные улучшения имеют прямую интерпретацию в терминах бизнес-метрик. Рост nDCG@10 коррелирует с увеличением кликабельности предложенных аналогов, а высокая полнота на уровне товаров повышает вероятность успешного завершения пользовательского сценария, включая конверсию в покупку. Таким образом, графовые признаки выступают не только как средство повышения офлайн-метрик, но и как практический инструмент улучшения пользовательского опыта и коммерческой эффективности системы.

В качестве направлений для дальнейших исследований представляют интерес: учёт временной динамики каталога и цен, расширение графовой модели за счёт мультимодальных сигналов, а также онлайн-оценка селективных стратегий ранжирования в A/B-экспериментах с фокусом на долгосрочные метрики удержания и жизненной ценности пользователя.

Список литературы

1. Theoretical understandings of product embedding for e-commerce machine learning / D. Xu [и др.] // Proceedings of the 14th ACM international conference on web search and data mining. — 2021. — С. 256—264.
2. *Краснов Ф. В.* Пороговые показатели полноты и точности для оценки системы извлечения информации о товарах на основе эмбеддингов // Бизнес-информатика. — 2024. — Т. 18, № 2. — С. 22—34.
3. *Li L.*, *Chen J.*, *Raghunathan S.* Recommender system rethink: Implications for an electronic marketplace with competing manufacturers // Information Systems Research. — 2018. — Т. 29, № 4. — С. 1003—1023.
4. *Yang D.-H.*, *Gao X.* Online retailer recommender systems: A competitive analysis // International Journal of Production Research. — 2017. — Т. 55, № 14. — С. 4089—4109.
5. *Krasnov F.*, *Kurushin F.* Reducing the long tail effect in e-commerce through self-attention // 36th Conference of Open Innovations Association FRUCT. Т. 36. — 2024. — С. 24—37.
6. *Brizan D. G.*, *Tansel A. U.* A. survey of entity resolution and record linkage methodologies // Communications of the IIMA. — 2006. — Т. 6, № 3. — С. 5.

7. An overview of end-to-end entity resolution for big data / V. Christophides [и др.] // ACM Computing Surveys (CSUR). — 2020. — Т. 53, № 6. — С. 1—42.
8. Synthesizing entity matching rules by examples / R. Singh [и др.] // Proceedings of the VLDB Endowment. — 2017. — Т. 11, № 2. — С. 189—202.
9. *Barlaug N.*, *Gulla J. A.* Neural networks for entity matching: A survey // ACM Transactions on Knowledge Discovery from Data (TKDD). — 2021. — Т. 15, № 3. — С. 1—37.
10. *Tuteja S.*, *Kumar R.* A unification of heterogeneous data sources into a graph model in e-commerce // Data Science and Engineering. — 2022. — Т. 7, № 1. — С. 57—70.
11. Semantic product search for matching structured product catalogs in e-commerce / J. I. Choi [и др.] // arXiv preprint arXiv:2008.08180. — 2020.
12. *Li M.*, *Yang C.-M.* E-Commerce User Shopping Preference Ranking Toward Million-Scale Products: A Hierarchical Feature Learning and Huge Purchase Graph Clustering Framework // IEEE Access. — 2025.
13. Deep graph embedding for ranking optimization in e-commerce / C. Chu [и др.] // Proceedings of the 27th ACM International Conference on Information and Knowledge Management. — 2018. — С. 2007—2015.
14. *Prajapat R.* Ranking: Science of Sorting in Ecommerce // AI-Powered Ecommerce: How Machine Learning Is Transforming Online Shopping. — Springer, 2024. — С. 149—175.
15. E-commerce search via content collaborative graph neural network / G. Xv [и др.] // Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining. — 2023. — С. 2885—2897.
16. Learning to rank using gradient descent / C. Burges [и др.] // Proceedings of the 22nd international conference on Machine learning. — 2005. — С. 89—96.
17. On the Foundations of Noise-free Selective Classification. / R. El-Yaniv [и др.] // Journal of Machine Learning Research. — 2010. — Т. 11, № 5.
18. *Geifman Y.*, *El-Yaniv R.* Selective classification for deep neural networks // Advances in neural information processing systems. — 2017. — Т. 30.
19. A path-constrained framework for discriminating substitutable and complementary products in e-commerce / Z. Wang [и др.] // Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. — 2018. — С. 619—627.

20. *Cho Y. C.*, *Ha J.* Consumer choice behavior on the web: The effects of product attributes on willingness to purchase // Journal of Business & Economics Research. — 2004. — Т. 2, № 10. — С. 75—87.

21. *Hartstein M.*, *Giannatou E.*, *Tegner M.* An Analysis of Learned Product Embeddings in an E-Commerce Context // Proceedings of the Nineteenth ACM Conference on Recommender Systems. — 2025. — С. 911—914.

22. *Huang Z.* Graph-based analysis for e-commerce recommendation. — 2005.

23. Hierarchical bipartite graph neural networks: Towards large-scale e-commerce applications / Z. Li [и др.] // 2020 IEEE 36th International Conference on Data Engineering (ICDE). — IEEE. 2020. — С. 1677—1688.

24. *Singh M.* Scalability and sparsity issues in recommender datasets: a survey // Knowledge and Information Systems. — 2020. — Т. 62, № 1. — С. 1—43.

25. *Idrissi N.*, *Zellou A.* A systematic literature review of sparsity issues in recommender systems // Social Network Analysis and Mining. — 2020. — Т. 10, № 1. — С. 15.

26. Data scarcity in recommendation systems: A survey / Z. Chen [и др.] // ACM Transactions on Recommender Systems. — 2025. — Т. 3, № 3. — С. 1—31.