DOI: 10.17323/2587-814X.2024.2.22.34

Пороговые показатели полноты и точности для оценки системы извлечения информации о товарах на основе эмбеддингов

Ф.В. Краснов
E-mail: krasnov.fedor2@wb.ru

Исследовательский центр ООО «ВБ СК» на базе Инновационного центра «Сколково», Москва, Россия

Аннотация

Современные системы извлечения информации о товарах для семантического поиска становятся все более сложными за счет использования дополнительных модальностей представления товаров, таких как пользовательское поведение, семантика языка и изображения. Однако добавление новой информации и усложнение моделей машинного обучения не обязательно ведут к улучшению показателей поиска, так как после извлечения производится ранжирование списка товаров, вносящее свое смещение. Тем не менее, бизнес-показатели продуктового поиска с ранжированием неполного списка товаров всегда будут хуже по сравнению с использованием полного списка, а от идеальной сортировки не соответствующих поисковому запросу товаров релевантность поисковой выдачи не улучшится. Поэтому основными показателями качества поиска для фазы извлечения товаров остаются полнота и точность по порогу k. В работе сопоставлено несколько архитектур систем извлечения товаров для семантического продуктового поиска на электронных торговых интернет-площадках. Для этого исследованы понятия пороговой полноты и точности для информационного поиска и выявлена зависимость этих показателей от порядка поисковой выдачи. Разработана автоматическая процедура расчета пороговой полноты и точности, позволяющая сравнивать эффективность систем извлечения информации. Предложенная автоматическая процедура протестирована на публичном наборе данных WANDS для нескольких ключевых архитектур. Полученные показатели полноты $R@1000 = 84\% \pm 9\%$ и точности $P@10 = 67\% \pm 17\%$ находятся на уровне SOTA моделей.

Ключевые слова: методы извлечения на основе эмбеддингов, информационный поиск, пороговые показатели, семантический поиск

Цитирование: 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. DOI: 10.17323/2587-814X.2024.2.22.34

Введение

Эффективность продуктового поиска критически важна для бизнеса электронных торговых интернет-площадок [1], поскольку согласно исследованию [2] более 90% пользователей принимают решение о покупке товара после использования поиска. Одна из самых ранних версий поисковых технологий Amazon обеспечила более 35% продаж [3]. Современный подход к продуктовому поиску [4–6] построен на парадигме информационного поиска (information retrieval), состоящей из двух фаз: извлечения и ранжирования документов. На электронной торговой интернет-площадке документами считаются данные о товарах или карточки товаров. Извлечение данных о товарах лежит в основе продуктового поиска. Если товар не найден, он не появится в поисковой выдаче и не будет отсортирован в порядке приоритетов покупателя, продавца и самой торговой интернет-площадки. Карточки товаров – это мультимодальные документы, поскольку данные о товаре могут быть представлены в виде текстового названия, списков характеристик, графических изображений, видео и отзывов покупателей. Извлечение данных необходимо делать из каждой модальности для наибольшей полноты. Комплексирование извлеченных из разных модальностей карточек товаров в единый список для дальнейшего ранжирования – отдельная задача, не рассматриваемая в настоящем исследовании. Современные подходы к извлечению на основе представлений в векторных пространствах высокой размерности с помощью искусственных нейронных сетей с глубоким обучением могут создавать единое пространство для объединения карточек товаров с разной модальностью.

Между лексическими методами извлечения [7] и методами извлечения на основе эмбеддингов (embedding based retrieval) существует принципиальное различие. Лексические методы извлечения основаны на наличии или отсутствии в документе определенного токена, поэтому про любой документ можно однозначно сказать, соответствует он поисковому запросу или нет. К примеру, существует поисковый запрос «юбка» и каталог с двумя карточками товаров (КТ1 и КТ2) с одной модальностью – название товара: КТ1 «юбка белая», КТ2 «брюки». С помощью лексических методов извлечения, примененных к подобному каталогу товаров, будет получено КТ1 и не получено КТ2. С применением методов извлечения на основе эмбеддингов КТ1 будет соответствовать поисковому запросу на 0,9, а КТ2 – на 0,1. Таким образом, лексические методы извлечения приводят к разряженному извлечению (sparse retrieval), а методы извлечения на основе эмбеддингов – к плотному извлечению (dense retrieval) карточек товаров. Применяя метод извлечения на основе эмбеддингов, необходимо определить порог релевантности для отсечения оптимального количества карточек товара. В случае высокого порога отсечения выдача становится короче и ее легче ранжировать, но появляется риск падения полноты выдачи. В случае низкого порога отсечения полнота будет высокой, но для ранжирования выдачи потребуются значительные вычислительные ресурсы. Это недопустимо, так как ранжирование производится в режиме, близком к реальному времени из-за необходимости учета различных факторов: локации пользователя, наличия товаров на складе, ценообразования. Таким образом, задача поиска оптимальных параметров системы извлечения крайне актуальна.

Бизнес-показатели систем продуктового поиска можно разделить на две категории: онлайн-показатели и автономные показатели. Онлайн-показатели измеряются во время фактического использования системы продуктового поиска под реальной нагрузкой. Они учитывают взаимодействие с пользователем, например, кликнул ли пользователь на найденную карточку товара или нет. Существует множество онлайн-показателей, но все они относятся к той или иной форме взаимодействия с

пользователем и не являются предметом настоящего исследования.

Автономные показатели измеряются в изолированной среде перед развертыванием новой версии системы информационного поиска. Они определяют, возвращается ли определенный набор релевантных результатов при извлечении документов с помощью системы. В научных статьях выделяют два типа автономных показателей: с учетом порядка и без учета порядка набора [7, 9]. К показателям, учитывающим порядок, относятся дисконтированный кумулятивный выигрыш (discounted cumulative gain, DCG), нормализованный дисконтированный кумулятивный выигрыш (normalized discounted cumulative gain, NDCG), среднеобратный ранг (mean reciprocal rank, MRR). К показателям, не учитывающим порядок относятся полнота и точность: именно они являются наиболее прозрачными мерами готовности системы к внедрению.

Разработка новых версий систем продуктового поиска – длительный и дорогостоящий комплекс организационно-технических мероприятий, критически важный для бизнеса. Оценка эффективности новой версии системы продуктового поиска – необходимый этап, который может выполняться неоднократно. Наличие прозрачных, информативных, не требующих больших затрат и научно обоснованных показателей повышает вероятность успеха внедрение новых версий систем продуктового поиска. Поэтому данное исследование сфокусировано именно на показателях полноты и точности.

Данная статья включает описание методики исследования, экспериментальные результаты, а также заключение.

1. Методика

Показатели оценки качества поиска нуждаются в точных трактовках для интерпретации результатов исследований. Наличие математической формулы показателя без детального описания может быть интерпретировано неоднозначно. Например, в книге [10] и в исследовании [11] приводится формула показателя точности для одного поискового запроса, хотя очевидно, что показатель точности варьируется для разных поисковых запросов. Приведение зависимостей показателей полноты и точности в исследованиях [5, 12] сделано без указания порога, что значительно затрудняет интерпретацию результатов. Выбор показателей для оценки качества продуктового поиска должен быть обоснован. Однако в научных статьях редко можно встретить обоснование использования тех или иных показателей. Например, в [12] для двух наборов данных выбраны разные показатели точности, для MS MARCO Dev [13] – *MRR@10*, а для TREC2019 DL [14] – *MAP@10*. Без внимания оставлено понимание алгоритма релевантности в формулах для *AP@1* в работе [15]. А в статье [16] рассмотрен показатель точности для высоких значений порога $k > 1000$, что требует отдельных обоснований, поскольку показатель точности важен для результата «первой страницы» поисковой выдачи. В исследовании [17] приведен только показатель полноты по порогам 10, 50, 100 без анализа показателя точности. Отметим, что формулы для показателей полноты и точности в статистике отличаются от формул для информационного поиска. Поэтому введем текстовое описание алгоритмов вычисления значений *R@k* и *P@k* для информационного поиска.

Определение 1: Пороговая полнота *R@k* – это среднее значение по всем поисковым запросам q_i, $Q = \{q_i\}$; для каждого поискового запроса q_i вычисляется пересечение множества всех карточек товаров, соответствующих поисковому запросу $C^g_{q_i}$ – истинной выдачи, с k первыми карточками товаров из упорядоченного по убыванию ранга списка извлеченных карточек товаров $C^r_{q_i}@k : \left|C^g_{q_i} \cap C^r_{q_i}@k\right|$, деленное на $\left|C^g_{q_i}\right|$ – количество всех карточек товаров, отвечающих поисковому запросу (1).

$$R@k = \frac{1}{|Q|}\sum_{q_i \in Q} \frac{\left|C^g_{q_i} \cap C^r_{q_i}@k\right|}{\left|C^g_{q_i}\right|}, \quad (1)$$

где

$|Q|$ – количество рассматриваемых поисковых запросов;

k – порог отсечки поисковой выдачи;

q_i – поисковый запрос $q_i \in Q$;

$C^g_{q_i}$ – множество всех товаров, соответствующих поисковому запросу q_i;

$C^r_{q_i}$ – поисковая выдача, множество всех товаров, найденных по поисковому запросу q_i.

По аналогии с *R@k* формула для показателя точности *P@k* имеет следующий вид (2):

$$P@k = \frac{1}{|Q|}\sum_{q_i \in Q} \frac{\left|C^g_{q_i} \cap C^r_{q_i}@k\right|}{k}. \quad (2)$$

Отличие формул для *R@k* и *P@k* заключается в знаменателе: для *R@k* в знаменателе стоит коли-

чество всех карточек товаров, соответствующих поисковому запросу $|C_{q_i}^{g}|$, а для $P@k$ – количество извлеченных карточек товаров, ограниченное порогом $k = \left|C_{q_i}^{r}@k\right|$.

В соответствии с формулами (1, 2) можно сделать вывод о модельном поведении показателей $R@k$, $P@k$ в зависимости от порога k. Для показателя полнота $R@k$ предельные значения показаны в формулах (3):

$$\lim_{k\to 0} R@k = 0,$$
$$\lim_{k\to \infty} R@k = 1. \quad (3)$$

Для показателя точность $P@k$ предельные значения представлены в формулах (4):

$$\lim_{k\to 0} P@k = 1,$$
$$\lim_{k\to \infty} P@k = 0. \quad (4)$$

Чтобы продемонстрировать наличие зависимости определенных по формулам (1, 2) пороговых показателей полноты и точности от порядка поисковой выдачи, сформулирована лемма 1.

Лемма 1. Значения пороговых показателей для полноты и точности зависят от порядка выдачи.

Более детально исследованы отношения показателей полноты и точности к упорядоченности выдачи. На *рисунке 1* приведен пример расчета показателей.

Из *рисунка 2* следует, что при вычислении показателей полноты и точности порядок выдачи влияет на значения показателей. Например, если бы товар с *Id4* находился бы ближе, чем *порог 3*, к началу выдачи, то *Точность@3* составила бы 2/3, а *Полнота@3* равнялась бы 2/7. Но в границах порога ($k = 3$) позиция товара с *Id4* не повлияет на зна-

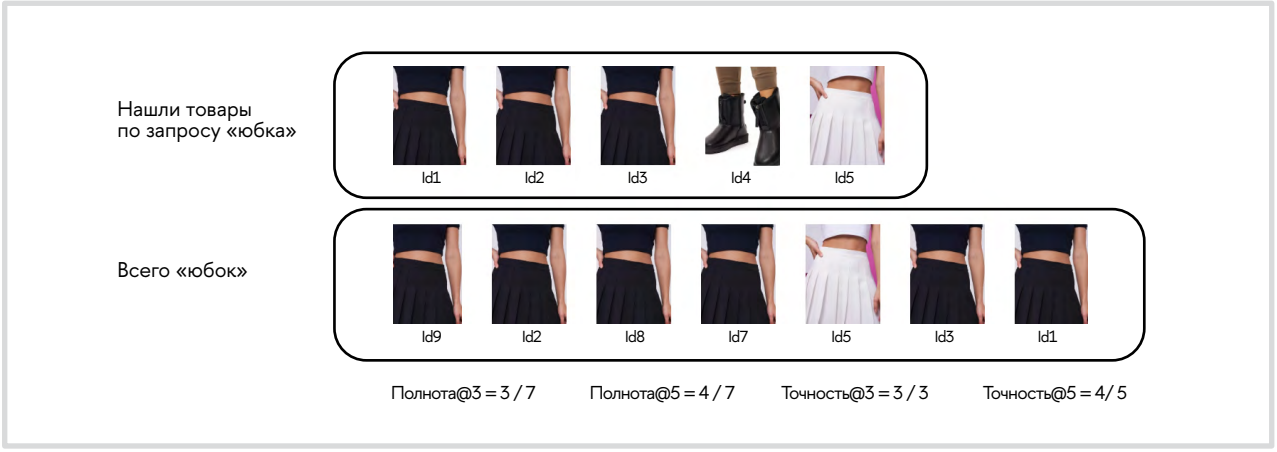


Рис. 1. Полнота и точность поисковой выдачи.

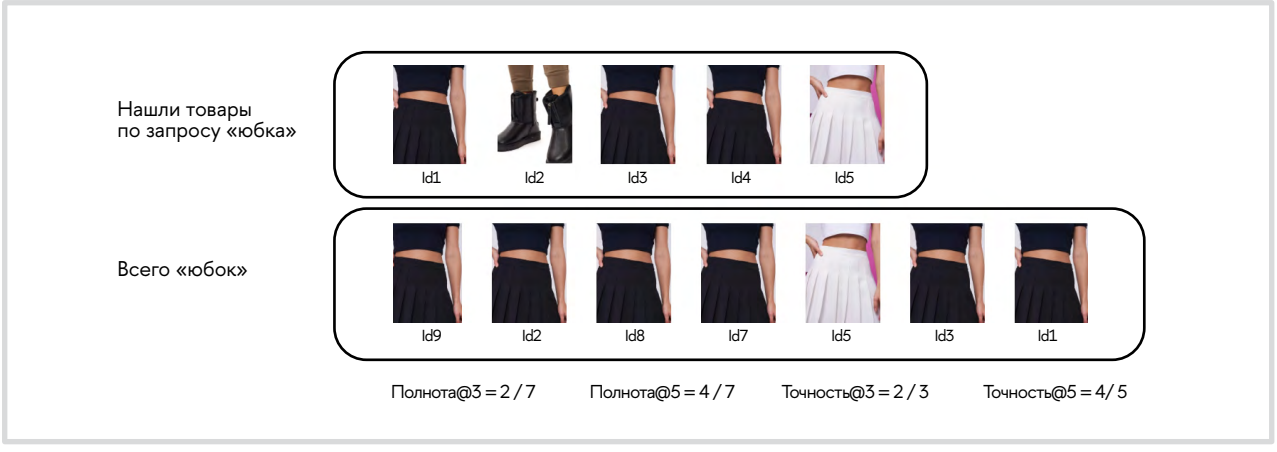


Рис. 2. Пороговая полнота и точность поисковой выдачи.

чения показателей по данному порогу (*рисунок 2*). Таким образом, пороговые (@k) показатели полноты и точности зависят от порядка выдачи. *Лемму 1* можно считать доказанной аналитически.

Учитывая, что пороговые параметры полноты и точности зависят от порядка товаров в выдаче, необходимо оценить систему извлечения в интегральном понимании, а не в одной точке k, при которой поведение показателя может быть смещенным. Поэтому целесообразно оценивать систему по сумме дискретных показателей точности для порогов k от 1 до K, где K – это один из гиперпараметров системы. Для примера, изображенного на *рисунке 1*, значение интегрального показателя пороговой точности:

$$AP@K = \frac{1}{K}\sum_{k=1}^{K} P@k.$$

Далее рассмотрено, как реализовано пересечение двух множеств товаров. На *рисунке 1* представлены Id_i товаров, тогда $C_{q_i}^g = [Id9, Id2, Id8, Id7, Id5, Id3, Id1]$, а $C_{q_i}^r@k = [Id1, Id2, Id3, Id4, Id5]$ для $k = 5$. Для вычисления компонента $\left|C_{q_i}^g \cap C_{q_i}^r@k\right|$ выполнена операция пересечения, при которой каждый элемент из $C_{q_i}^g$ сравнен с каждым элементом из $C_{q_i}^r@k$. В рассматриваемом случае элементами являются товары, обладающие несколькими модальностями, поэтому их можно сравнить с помощью различных алгоритмов. На *рисунке 1* изображены две модальности товара: цифровой идентификатор (Id) и изображение. Кроме этих модальностей в научной литературе рассматривают и другие модальности товара, основанные на его текстовых представлениях, например, по названию товара eBay [16], по характеристикам товара в исследовании Amazon [5]. В общем случае решение задачи об идентичности двух товаров может быть решено на основе функции суперпозиции сходств нескольких модальностей, например, как в исследовании [18].

Общий вид функции идентичности товаров обозначен как $S_V(\cdot,\cdot)$, где V – это векторное пространство, в котором будут представлены товары для сравнения, а $\cdot$ – модальность товара. Далее проанализированы следующие векторные пространства: N – пространство натуральных чисел, цифровых идентификаторов товаров, T – пространство строк, названий карточек товаров, I – пространство растровых изображений товаров. Тогда можно записать, что $V \in \{N, T, I\}$. Описано каждое из векторных пространств $\{N, T, I\}$ с позиции разряженности и плотности (sparse/dense).

- Пространство N постулирует, что карточки идентичны в единственном случае, когда равны их цифровые идентификаторы. Цифровой идентификатор представляет собой уникальный номер карточки товара. У каждой карточки может быть только одна идентичная ей карточка товара с таким же цифровым идентификатором. Пространство N является разряженным (sparse), так как отношение количества пар идентичных товаров ко всем возможным парам товаров близко к нулю. Другими словами, если рассматривать матрицу идентичности товаров, то тождественные пары товаров будут располагаться на диагонали.
- Пространство T постулирует, что товары с идентичными названиями являются дублями (идентичными). Это значительно более мягкое условие идентичности товаров, чем в пространстве N. Так, все товары с названием «синий лак для ногтей» будут идентичными. Очевидно, что таких товаров в каталоге значительно больше, чем один. В пространстве T можно ввести дополнительные подпространства, позволяющие сравнивать карточки товаров между собой еще более интуитивно. Например, возможно учитывать только наличие токенов, но не их порядок (Bag of Words, BoW). Такое подпространство для сравнения будет постулировать идентичными карточками товаров с названиями «лак для ногтей синий» и «синий лак для ногтей». Пространство T также разряженное (sparse).
- Пространство I определяет идентичность товаров через сравнение их изображений. На *рисунке 1* приведено изображение товара «юбка», имеющее разные цифровые идентификаторы, но в пространстве I такие товары будут идентичными. Сравнение изображений алгоритмически рассмотрено как приведение изображений в компактное пространство меньшей размерности (embedding) и вычисление косинусной близости. Такой подход добавляет еще один гиперпараметр – порог косинусной близости изображений товаров, определяющий, что товары идентичны. Пространство I является плотным (dense), другими словами, все элементы пространства T идентичны в той или иной степени.

Таким образом, необходимо определять векторные пространства $\{N, T, I\}$ для функции идентичности товаров $S_V(\cdot,\cdot)$.

Проанализировано, какие гиперпараметры управляют семантическим продуктовым поиском. Выше рассмотрены возможные пространства для функции идентичности $S_V(\cdot,\cdot)$, но не для самой системы извлечения информации о товарах. Современные системы извлечения информации о товарах на основе эмбеддингов строятся на композиции моделирования семантики языка, изображения товаров и поведения пользователей (*рис. 3*).

Композиционный подход к извлечению описан в исследовании Amazon [5], в нем показано, что различные типы поведения пользователей могут при комбинировании выдач привести к значительному улучшению показателей. В исследовании системы извлечения для продуктового поиска на интернет-площадке Taobao [6] представлена модель под названием «Многоуровневый глубокий семантический поиск продуктов» (multi-grained deep semantic product retrieval, MGDSPR) для одновременного моделирования семантики запросов и исторических данных о поведении пользователей с целью получения более полной выдачи товаров с хорошей релевантностью. На торговой интернет-площадке Walmart также используют комбинацию источников для системы извлечения [19]: архитектура семантической модели представляет собой структуру из двух «башен», каждая «башня» – это искусственная нейронная сеть глубокого обучения, формирующая в векторном пространстве компактное представление для поискового запроса и продукта соответственно. Оценка пары «запрос – продукт» реализована с помощью функции потерь на основании косинусной близости. Пример исследования системы извлечения в разряженном (sparse) пространстве [20] от исследователей компании Tencent демонстрирует улучшение эффективности по показателю «полнота» при уменьшении потребляемого места на диске. Для семантического поиска продуктов в компании Etsy, согласно исследованию [21], также используется мультимодальность продуктов в своей модели UPPER, но немного шире обычного, так как персонализация трактуется обучение «двух башенной» модели на поведенческих данных пользователей.

Несмотря на то, что в качестве показателей эффективности в исследованиях [5, 6, 19] использованы пороговые показатели полноты и точности в различных вариациях, необходимость ранжировать комбинированную выдачу оставлена без должного внимания. Вторая общая особенность исследований [5, 6, 20, 21] и многих других – использование пороговых показателей полноты как прокси-показателей модели, а не как часть функции потерь для поиска оптимальных параметров системы извлечения. Кроме того, следует отметить необходимость отдельного анализа формул для показателей пороговой полноты и точности, которые не приведены в отдельных исследованиях [6, 19]. В-третьих, в статьях [5, 6, 20, 21] не приводится ошибка, возникающая при вычислении пороговых показателей полноты и точности по набору поисковых запросов.

В *таблице 1* представлены пороговые показатели полноты *R@1000* из приведенных выше исследований.

Нецелесообразно считать показатель полноты выдачи для отраслевых систем извлечения для значений $k < 100$, так как знаменатель в формуле (1) будет иметь слишком большие значения. Однако показатель пороговой точности для $k = 10$, наоборот, хорошо отражает качество системы извлечения, так как соответствует наиболее просматрива-

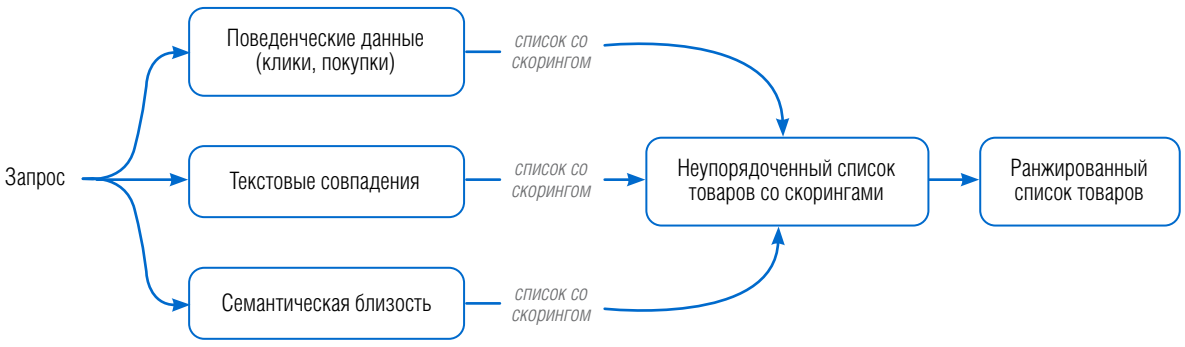


Рис. 3. Композиция выдач и ранжирование.

Таблица 1.

Показатели отраслевых исследований

Модель	Показатель	Значение
UPPER [21]	*R@1000*	0,85
MGDSPR [6]	*R@1000*	0,85
SPS [5]	*R@1000*	0,79
SPS [5]	*MAP@10*	0,74

емым позициям товаров. Поэтому тот факт, что в исследованиях [5, 6, 20, 21] не производилось измерение показателя пороговой точности, удивляет.

На основании доказанной леммы и результатов современных исследований лидеров индустрии сформулирована основная задача исследования, которая состоит в том, чтобы провести экспериментальную апробацию автоматизированной процедуры сравнения разных версий систем извлечения информации о товарах для продуктового поиска на основе автономных показателей пороговой полноты и точности.

2. Эксперимент

Для решения исследовательской задачи проведен цифровой эксперимент: выбрать размеченный вручную набор данных D_G; обучить три модели системы извлечения: DE («двухбашенная» модель с одной модальностью), DE2 («двухбашенная» модель с двумя модальностями), модель «двух башен» с одним энкодером – Single Encoder с одной модальностью; сравнить результаты систем извлечения на наборе данных по показателям пороговой полноты и точности, принятым в отрасли (benchmark).

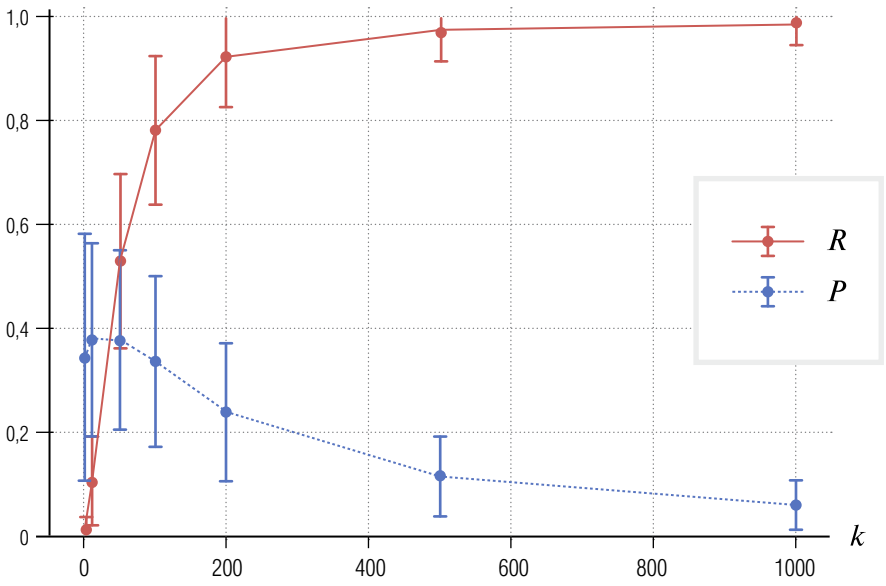
В качестве набора данных D_G выбран WANDS [22], позволяющий проводить объективный бенчмаркинг и оценку поисковых систем на основе набора данных электронной коммерции. Его ключевые характеристики включают:

- 42 994 товара-кандидата;
- 480 запросов;
- 233 448 оценок релевантности (запроса, товар).

Набор данных WANDS обладает трехуровневой разметкой пар «запрос – товар»: «полностью соответствует» (Exact), «частично соответствует» (Partial), «не соответствует» (Irrelevant). Поэтому для обучения моделей системы извлечения использованы только два значения для построения функции потерь: 1 для Exact и –1 для Irrelevant. Полученные классы сбалансированы при обучении.

На *рисунке 4* приведены зависимости пороговой полноты и точности на основе разметки D_G, построенные по формулам (1) и (2) соответственно.

В зависимости от пороговой точности (P) на *рисунке 4* отражены достаточно сильные отклоне-



ния от модельного поведения (4). При значениях порога $k = 10, 50$ пороговая точность составляет $P@10 = 0{,}37 \pm 0{,}17$, $P@50 = 0{,}38 \pm 0{,}18$. Для понимания причин, влияющих на допустимый интервал значений, на *рисунке 5* представлены зависимости полноты и точности для отдельных запросов.

Для эксперимента выбраны три архитектуры системы извлечения – DE (*рис. 6*), DE2 (*рис. 7*), SE (*рис. 8*).

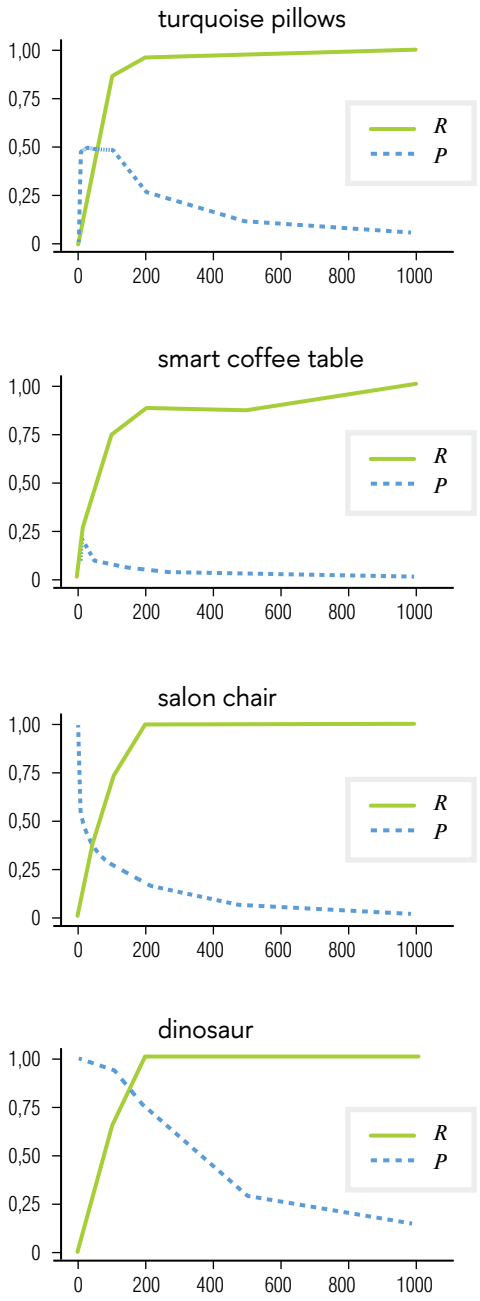


Рис. 5. Полнота и точность для отдельных запросов без системы извлечения.

Рис. 6. Модель извлечения DE.

Для обучения моделей систем извлечения использован набор данных со следующими параметрами: оптимизатор AdamW, скорость обучения циклически менялась от значения 0,01 до 0,1, 500 эпох с ранней остановкой. В качестве токенизатора использован BPE метод с размером словаря 16 тысяч токенов для товаров, 512 токенов для запросов для моделей DE и DE2. Для модели SE размер словаря составляет 16 тысяч токенов. Среди гиперпараметров модели извлечения акцентировано внимание на размерности таблицы векторов токенов, влияющей как на скорость прогнозирования, так и на размер модели в памяти. В рамках эксперимента зафиксировано следующее явление: при переходе от размерности 256 к 32 ошибка валидации увеличивается на 3,5% и, хотя размер таблицы уменьшается в 8 раз, скорость прогнозирования и обучения возрастают более чем в 3 раза. Все зависимости валидационных ошибок от размерности векторов токенов представлены на *рисунке 9*.

К набору данных D_G применены три выбранных системы извлечения и получены товары-кандидаты, по которым определены пороговые показатели полноты и точности. В результате измерения показателей различных систем извлечения получены следующие результаты (*таблица 2*).

В *таблице 2* в строке «Разметка» приведены значения $R@1000$ и $P@10$ для набора данных D_G без применения систем извлечения. Точность при пороге $k = 10$ без применения систем извлечения са-

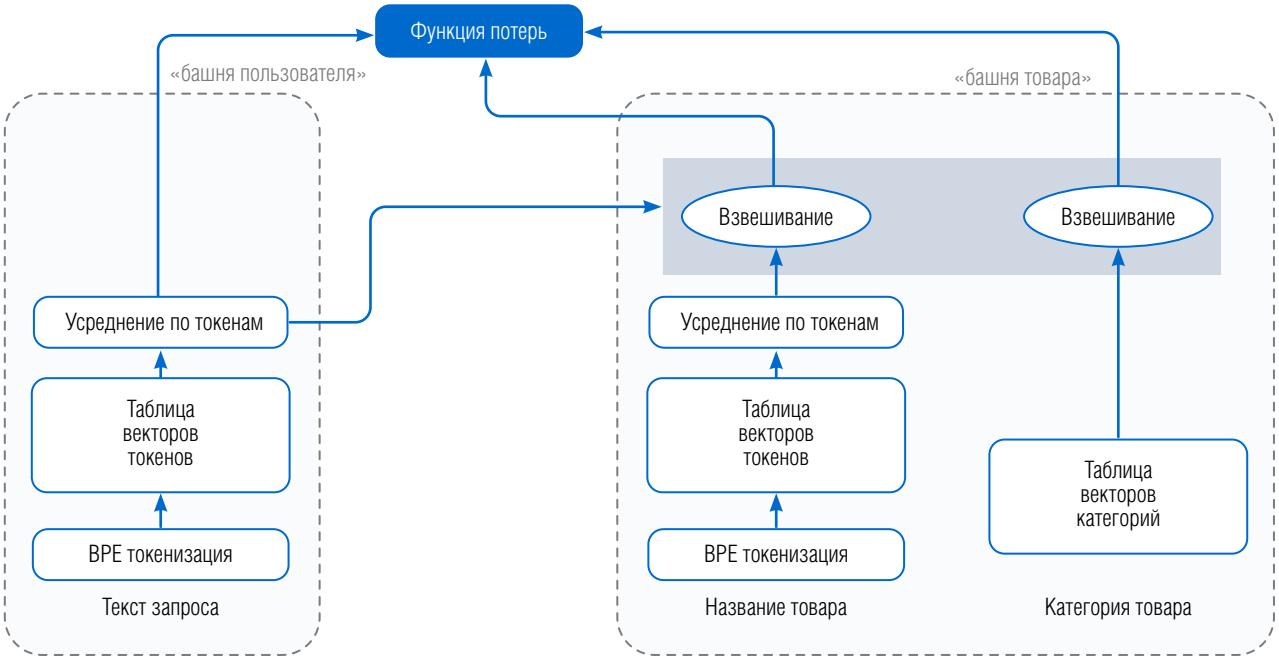


Рис. 7. Модель извлечения DE2.

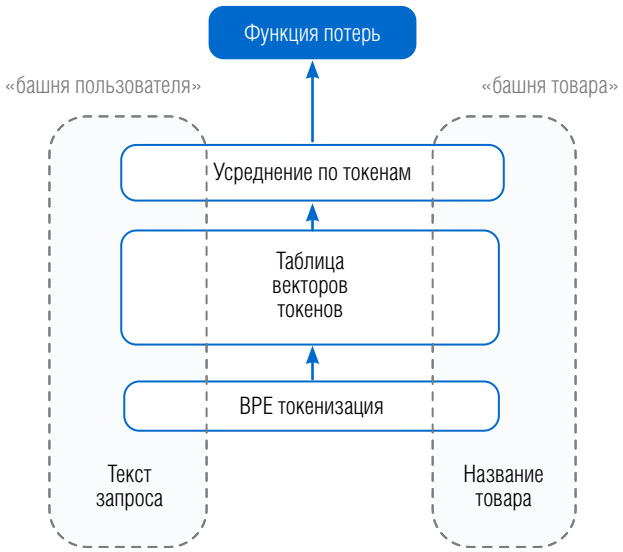


Рис. 8. Модель извлечения SE.

мая низкая из всех рассматриваемых моделей – 0,37. Это означает, что при разметке последовательности примеров не уделили внимания. С помощью моделей извлечения получилось осуществить сортировку примеров в порядке убывания показателя точности так, чтобы пороговое значение точности стало выше у модели DE – 0,68. Показательно,

Таблица 2.

Значения пороговых показателей различных систем извлечения

Модель	*R@1000*		*P@10*	
	mean	std	mean	std
DE	0,75	0,10	0,68	0,16
DE2	0,73	0,11	0,66	0,17
SE	0,84	0,09	0,67	0,17
Разметка	0,99	0,03	0,37	0,17

полнота при пороге $k = 1000$ без применения систем извлечения самая высокая из всех рассматриваемых моделей, что является самопроверкой для кода вычислений. Ошибки (std) для всех моделей принимают близкие значения для точности 0,10, 0,11, 0,09, для полноты – 0,16, 0,17, 0,17. Следовательно, модели «ошибаются» на разных запросах однотипно. Модель DE2 не показала лучших результатов, хотя при обучении была задействована дополнительная модальность. Модель SE с наименьшим количеством параметров для обучения показала лучшую полноту $0,84 \pm 0,09$.

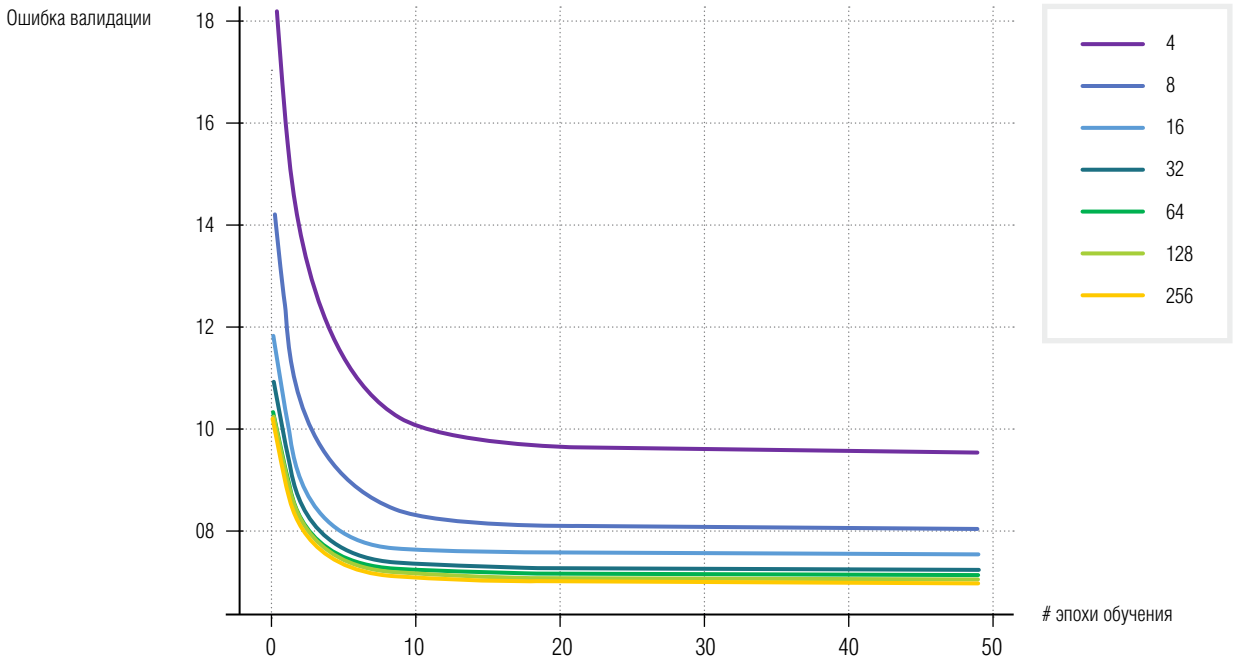


Рис. 9. Зависимости валидационных ошибок от размерности векторов токенов.

Заключение

Оценка систем продуктового поиска имеет решающее значение для принятия обоснованных бизнес-решений электронными торговыми интернет-площадками. Многие крупные технологические компании добиваются успеха благодаря качественно построенному продуктовому поиску. Важно, что измерение эффективности продуктового поиска – это непрерывный процесс, связанный с постоянным улучшением данных и научных достижений в области машинного обучения. Пороговые показатели полноты и точности обладают высокой интерпретируемостью в отличии от других показателей эффективности продуктового поиска, могут служить объективными мерами качества как разметки данных, так и работы систем извлечения информации о товарах. На публичном наборе данных WANDS продемонстрировано, что относительно простые архитектуры моделей систем извлечения могут достигать значений показателей из статей лидеров отрасли со значительно превосходящим числом параметров моделей. В рамках исследования разработана автоматизированная процедура для расчета пороговых показателей применительно к набору поисковых запросов. В результате данного исследования создана и экспериментально опробована автоматизированная процедура измерения эффекта для систем извлечения информации о товарах (first stage retrieval). Перспектива исследования – проведение эксперимента на большем количестве модальностей карточек товаров, измерения эффекта от предобученных моделей по сравнению с обучением «с нуля» и дообучением моделей извлечения информации о товарах. ■

Литература

1. Матвеев М.Г., Алейникова Н.А., Титова М.Д. Технология поддержки принятия решений продавца на маркетплейс в условиях конкуренции // Бизнес-информатика. 2023. Т. 17. № 2. С. 41–54. https://doi.org/10.17323/2587-814X.2023.2.41.54

2. Luo C., Goutam R., Zhang H., Zhang C., Song Y., Yin B. Implicit query parsing at Amazon product search // Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023. P. 3380–3384. https://doi.org/10.1145/3539618.3591858

3. Linden G., Smith B., York J. Amazon.com recommendations: Item-to-item collaborative filtering // IEEE Internet computing. 2003. Vol. 7. No. 1. P. 76–80.

4. Huang P., He X., Gao J., Deng L., Acero A., Heck L. Learning deep structured semantic models for web search using clickthrough data // Proceedings of the 22nd ACM international conference on Information Knowledge Management. 2013. P. 2333–2338. https://doi.org/10.1145/2505515.2505665

5. Nigam P. Semantic product search // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019. P. 2876–2885. https://doi.org/10.1145/3292500.3330759

6. Li S. Embedding-based product retrieval in Taobao search // Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021. P. 3181–3189. https://doi.org/10.1145/3447548.3467101

7. Краснов Ф.В., Смазневич И.С., Баскакова Е.Н. Проблема потери решений в задаче поиска схожих документов: Применение терминологии при построении векторной модели корпуса // Бизнес-информатика. 2021. Т. 15. № 2. С. 60–74. https://doi.org/10.17323/2587-814X.2021.2.60.74

8. Mitra B., Craswell N. Neural models for information retrieval // arXiv:1705.01509. 2017. https://doi.org/10.48550/arXiv.1705.01509

9. Gudivada V.N., Rao D.L., Gudivada A.R. Information retrieval: concepts, models, and systems // Handbook of statistics. 2018. Vol. 38. P. 331–401. https://doi.org//10.1016/bs.host.2018.07.009

10. Buttcher S., Clarke C.L.A., Cormack G.V. Information retrieval: Implementing and evaluating search engines. The MIT Press: Cambridge, Massachusetts, London, England, 2016.

11. Leonhardt L.J. Efficient and Explainable Neural Ranking. PhD thesis. Hannover: Gottfried Wilhelm Leibniz Universität, 2023. https://doi.org/10.15488/15769

12. Campos D.F., Nguyen T., Rosenberg M., Song X., Gao J., Tiwary S., Majumder R., Deng L., Mitra B. MS MARCO: A human generated MAchine Reading COmprehension dataset // arXiv: 1611.09268. 2016. https://doi.org/10.48550/arXiv.1611.09268

13. Craswell N., Mitra B., Yilmaz E., Campos D., Voorhees E.M. Overview of the TREC 2019 deep learning track // arXiv: 2003.07820. 2020. https://doi.org/10.48550/arXiv.2003.07820

14. Leonhardt J., Müller H., Rudra K., Khosla M., Anand A., Anand A. Efficient neural ranking using forward indexes and lightweight encoders // ACM Transactions on Information Systems. 2023. https://doi.org/10.1145/3631939

15. Gao L., Dai Z., Chen T., Fan Z., Durme B.V., Callan J. Complement lexical retrieval model with semantic residual embeddings // Advances in Information Retrieval. ECIR 2021. Lecture Notes in Computer Science. 2021. Vol. 12656. P. 146–160. https://doi.org/10.1007/978-3-030-72113-8_10

16. Trotman A., Degenhardt J., Kallumadi S. (2017) The architecture of eBay search // SIGIR Workshop on eCommerce. eCOM@ SIGIR. Tokyo, Japan, 2017.

17. Chang W., Jiang D., Yu H., Teo C.H., Zhang J., Zhong K., Kolluri K., Hu Q., Shandilya N., Ievgrafov V., Singh J., Dhillon I.S. Extreme multi-label learning for semantic matching in product search // Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021. P. 2643–2651. https://doi.org/10.1145/3447548.3467092

18. Краснов Ф.В. Оценка временной сложности для задачи поиска идентичных товаров для электронной торговой площадки на основании композиции моделей машинного обучения // International Journal of Open Information Technologies. 2023. Vol. 11. No. 2. P. 72–76.

19. Magnani A., Liu F., Chaidaroon S., Yadav S., Suram P.R., Puthenputhussery A., Chen S., Xie M., Kashi A., Lee T., Liao C. Semantic retrieval at Walmart // Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022. P. 3495–3503. https://doi.org/10.1145/3534678.3539164

20. Gan Y., Ge Y., Zhou C., Su S., Xu Z., Xu X., Hui Q., Chen X., Wang Y., Shan Y. Binary embedding-based retrieval at Tencent // Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023. P. 4056–4067. https://doi.org/10.1145/3580305.3599782

21. Jha R., Subramaniyam S., Benjamin E., Taula T. Unified embedding based personalized retrieval in Esty search // arXiv: 2306.04833. 2023. https://doi.org/10.48550/arXiv.2306.04833

22. Chen Y., Liu S., Liu Z., Sun W., Baltrunas L., Schroeder B. WANDS: Dataset for product search relevance assessment // Advances in Information Retrieval. ECIR 2022. Lecture Notes in Computer Science. Vol. 13185. P. 128–141. https://doi.org/10.1007/978-3-030-99736-6_9

Об авторах

Краснов Федор Владимирович

к.т.н.;

специалист по поисковым и рекомендательным системам в электронной коммерции, сотрудник Исследовательского центра ООО «ВБ СК» на базе Инновационного Центра Сколково, Россия, г. Москва, ул. Нобеля, 5;

E-mail: krasnov.fedor2@wb.ru

ORCID: 0000-0002-9881-7371

Embedding-based retrieval: measures of threshold recall and precision to evaluate product search

Fedor V. Krasnov
E-mail: krasnov.fedor2@wb.ru

Research Center of WB SK LLC, Moscow, Russia

Abstract

Modern product retrieval systems are becoming increasingly complex due to the use of extra product representations, such as user behavior, language semantics and product images. However, adding new information and complicating machine learning models does not necessarily lead to an improvement in online and business search performance, since after retrieval the product list is ranked, which introduces its own bias. Nevertheless, the business performance of a product search will be worse from ranking an incomplete list of products than a complete one, and the relevance of search results will not improve from perfect sorting of products that do not match the search query. Therefore, the main quality indicators for the products retrieval phase remain Recall and Precision at the k threshold. This paper compares several architectures of product retrieval systems in product search for e-commerce. To do this, the concepts of threshold Recall and Precision for information retrieval are investigated and the dependence of these measures on the order of issuance is revealed. An automatic procedure has been developed for calculating $R@k$ and $P@k$, which allows us to compare the effectiveness of information retrieval systems. The proposed automatic procedure has been tested on the WANDS public dataset for several key architectures. The obtained values $R@1000 = 84\% \pm 9\%$ and $P@10 = 67\% \pm 17\%$ are at the level of SOTA models.

Keywords: embedding-based retrieval, information retrieval, threshold metrics, semantic product search

Citation: Krasnov F.V. (2024) Embedding-based retrieval: measures of threshold recall and precision to evaluate product search. *Business Informatics*, vol. 18, no. 2, pp. 22–34. DOI: 10.17323/2587-814X.2024.2.22.34

References

1. Matveev M.G., Aleynikova N.A. Titova M.D. (2023) Decision support technology for a seller on a marketplace in a competitive environment. *Business Informatics*, vol. 17, no. 2, pp. 41–54. https://doi.org/10.17323/2587-814X.2023.2.41.54
2. Luo C., Goutam R., Zhang H., Zhang C., Song Y., Yin B. (2023) Implicit query parsing at Amazon product search. Proceedings of the *46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://doi.org/10.1145/3539618.3591858
3. Linden G., Smith B., York J. (2003) Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, vol. 7, no. 1, pp. 76–80.
4. Huang P., He X., Gao J., Deng L., Acero A., Heck L. (2013) Learning deep structured semantic models for web search using clickthrough data. Proceedings of the *22nd ACM international conference on Information Knowledge Management*. https://doi.org/10.1145/2505515.2505665
5. Nigam P., Song Y., Mohan V., Lakshman V., Ding W., Shingavi A., Teo C.H., Gu H., Yin B. (2019) Semantic Product Search. Proceedings of the *25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*. https://doi.org/10.1145/3292500.3330759
6. Li S., Lv F., Jin T., Lin G., Yang K., Zeng X., Wu X., Ma Q. (2021) Embedding-based product retrieval in Taobao search. Proceedings of the *27th ACM SIGKDD Conference on Knowledge Discovery Data Mining*. https://doi.org/10.1145/3447548.3467101

7. Krasnov F.V., Smaznevich I.S., Baskakova E.N. (2021) The problem of loss of solutions in the task of searching similar documents: Applying terminology in the construction of a corpus vector model. *Business Informatics*, vol. 15, no. 2, pp. 60–74. https://doi.org/10.17323/2587-814X.2021.2.60.74

8. Mitra B., Craswell N. (2017) Neural models for information retrieval. *arXiv*: 1705.01509. https://doi.org/10.48550/arXiv.1705.01509

9. Gudivada V.N., Rao D., Gudivada A.R. (2018) Information retrieval: concepts, models, and systems. *Handbook of Statistics*, vol. 38, pp. 331–401. https://doi.org//10.1016/bs.host.2018.07.009

10. Büttcher S., Clarke C.L.A., Cormack G.V. (2010) *Information retrieval: Implementing and evaluating search engines*. The MIT Press: Cambridge, Massachusetts, London, England.

11. Leonhardt J. (2023) *Efficient and explainable neural ranking*. PhD thesis. Hannover: Gottfried Wilhelm Leibniz Universität. https://doi.org/10.15488/15769

12. Campos D.F., Nguyen T., Rosenberg M., Song X., Gao J., Tiwary S., Majumder R., Deng L., Mitra B. (2016) MS MARCO: A human generated MAchine Reading COmprehension dataset. *arXiv*: 1611.09268. https://doi.org/10.48550/arXiv.1611.09268

13. Craswell N., Mitra B., Yilmaz E., Campos D., Voorhees E.M. (2020) Overview of the TREC 2019 deep learning track. *arXiv*: 2003.07820. https://doi.org/10.48550/arXiv.2003.07820

14. Leonhardt, J., Müller, H., Rudra, K., Khosla, M., Anand, A., Anand, A. (2023) Efficient neural ranking using forward indexes and lightweight encoders. *ACM Transactions on Information Systems*. https://doi.org/10.1145/3631939

15. Gao L., Dai Z., Chen T., Fan Z., Durme B.V., Callan J. (2021) Complement lexical retrieval model with semantic residual embeddings. *Advances in Information Retrieval. ECIR 2021. Lecture Notes in Computer Science*, vol. 12656, pp. 146–160. https://doi.org/10.1007/978-3-030-72113-8_10

16. Trotman A., Degenhardt J., Kallumadi S. (2017) The Architecture of eBay Search. *SIGIR Workshop on eCommerce. eCOM@ SIGIR. Tokyo, Japan, August 2017.*

17. Chang W., Jiang D., Yu H., Teo C.H., Zhang J., Zhong K., Kolluri K., Hu Q., Shandilya N., Ievgrafov V., Singh J., Dhillon I.S. (2021) Extreme multi-label learning for semantic matching in product search. Proceedings of the *27th ACM SIGKDD Conference on Knowledge Discovery Data Mining*. https://doi.org/10.1145/3447548.3467092

18. Krasnov F. (2023) Estimation of time complexity for the task of retrieval for identical products for an electronic trading platform based on the decomposition of machine learning models. *International Journal of Open Information Technologies*, vol. 11, no. 2, pp. 72–76.

19. Magnani A., Liu F., Chaidaroon S., Yadav S., Suram P.R., Puthenputhussery A., Chen S., Xie M., Kashi A., Lee T., Liao C. (2022) Semantic retrieval at Walmart. Proceedings of the *28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3534678.3539164

20. Gan Y., Ge Y., Zhou C., Su S., Xu Z., Xu X., Hui Q., Chen X., Wang Y., Shan Y. (2023) Binary embedding-based retrieval at Tencent. Proceedings of the *29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Long Beach CA USA, August 6–10, 2023*, pp. 4056–4067. https://doi.org/10.1145/3580305.3599782

21. Jha R., Subramaniyam S., Benjamin E., Taula T. (2023) Unified embedding based personalized retrieval in Esty search. *arXiv*: 2306.04833. https://doi.org/10.48550/arXiv.2306.04833

22. Chen Y., Liu S., Liu Z., Sun W., Baltrunas L., Schroeder B. (2022) WANDS: Dataset for product search relevance assessment. *Advances in Information Retrieval. ECIR 2022. Lecture Notes in Computer Science*, vol. 13185, pp. 128–141. https://doi.org/10.1007/978-3-030-99736-6_9

About the author

Fedor V. Krasnov

Cand. Sci. (Tech.);

Specialist in information retrieval and recommendation systems in e-commerce, Researcher of the Research Center of WB SK LLC based on the Skolkovo Innovation Center, 5, St. Nobel, Moscow, Russia;

E-mail: krasnov.fedor2@wb.ru

ORCID: 0000-0002-9881-7371