DOI: 10.15514/ISPRAS-2024-36(5)-16

H1: гибридная система извлечения информации для поиска товаров в электронной торговле

[1] *Ф.В. Краснов, ORCID: 0000-0002-9881-7371 <krasnov.fedor2@wb.ru>*

[1] *Исследовательский центр ООО "ВБ СК" на базе Инновационного Центра Сколково Россия, 121205, Москва, ул. Нобеля, д. 5.*

Аннотация. В данной статье обоснована продуктивность использования системы H1 для поиска товаров различных поставщиков на торговой интернет-площадке. Как и все современные системы поиска товаров, гибридная система H1 соединяет в себе преимущества лексических методов извлечения товаров и семантических методов, основанных на многомерных векторных представлениях. Новизна предложенного подхода заключается в объединении методов извлечения на уровне токенов. Дополнительное преимущество H1, по сравнению с другими индустриальными системами, – обработка поисковых запросов, состоящих из нескольких слов. Например, поисковые запросы «конфеты рот фронт», «gloria jeans детская одежда» будут выделять сущность бренда в отдельный токен – «рот фронт», «gloria jeans», что позволит уменьшить размер модели и улучшить автономные показатели системы извлечения. Полученные на публичном наборе данных WANDS значения показателей усредненной пороговой точности mAP@12 = 56.1% и пороговой полноты R@1k = 86.6% для H1 превышают самые современные аналоги.

Ключевые слова: информационный поиск по товарам; распознавание бизнес-терминов; суб-словарная токенизация; трансформеры; электронная коммерция.

Для цитирования: Краснов Ф.В. H1: гибридная система извлечения для поиска товаров в e-commerce. Труды ИСП РАН, том 36, вып. 5, 2024 г., стр. 227–240. DOI: 10.15514/ISPRAS–2024–36(5)–16.

H1: Hybrid Retrieval System for Product Search in E-Commerce

[1] *F.V. Krasnov,* ORCID: 0000-0002-9881-7371 <krasnov.fedor2@wb.ru>

[1] *Research Center of WB SK LLC based on the Skolkovo Innovation Center, 5, Nobel b., Moscow, 121205, Russia.*

Abstract. This paper presents the effectiveness of using the H1 system for retrieving products from various vendors in the marketplace. The H1 system is a hybrid model that combines the benefits of lexical-based and semantic-based retrieval techniques, similar to other state-of-the-art product retrieval systems. The novelty of this approach lies in its combination of token-level retrievals. The advantage of the H1 system over other existing solutions is its ability to handle complex search queries containing brands with multi-word brands. For example, search queries like "new balance sneakers" and "Gloria Jeans children's clothing" will be split into separate tokens "new balance" and "Gloria Jeans", respectively, which helps reduce the retrieval model's size and improves its autonomous performance. The H1 system achieved mAP@12 score of 56.1% and R@1K score of 86.6% on the public WANDS dataset, outperforming other state-of-the-art models. These results

demonstrate the effectiveness of the approach and its potential for improving product search experiences for online shoppers.

Keywords: product information search; entity recognition; Sentence Piece; transformers; ColBERT; e-commerce.

For citation: Krasnov F.V. H1: hybrid retrieval system for product search in e-commerce. Trudy ISP RAN/Proc. ISP RAS, vol. 36, issue 5, 2024. pp. 227-240 (in Russian). DOI: 10.15514/ISPRAS-2024-36(5)-16.

1. Введение

Каталог товаров интернет-площадки включает миллиарды товарных позиций, для его сканирования система поиска товаров должна отвечать критериям скорости, точности и полноты. Типичные средства поиска товаров на первом этапе используют модель поиска по набору слов, которая вычисляет оценку релевантности на основе эвристики, определенной для точного совпадения слов между поисковыми запросами и текстовыми представлениями товаров. Такие лексические модели, как BM25 [1], не устаревают на протяжении десятилетий и по-прежнему широко используются сегодня. Нейросетевые методы извлечения позволяют улучшить показатели эффективности поиска, но обладают своими недостатками [2-5]. Поэтому естественно желание исследователей объединить всё лучшее от лексических и нейросетевых методов в гибридную модель. Недостатки лексических моделей извлечения хорошо изучены: (E1) возможное несоответствие словаря запроса и документа [6-7] приводит к деградации полноты поиска; (E2) отсутствие смыслового понимания документа и запроса [8] уменьшает точность поиска. Из-за ограничений лексические модели не справляются с извлечением релевантных документов в информационном поиске. В течение последних десятилетий предпринимались постоянные попытки решения данных проблем путем расширения запросов [9-12], расширения документа [13-15], создания моделей зависимости токенов [16-17], тематических моделей [18-19], моделей машинного перевода для информационного поиска [20-21]. Однако прогресс в исследованиях лексических моделей извлечения относительно медленный, поскольку большинство из этих подходов все еще находится в рамках парадигмы дискретного, разряженного лексического представления и неизбежно наследует ее ограничения.

Наряду с развитием репрезентативных методов обучения в информационном поиске наблюдается рост исследовательского интереса к моделям семантического поиска на этапе извлечения информации о документах. Начиная с 2013 года развитие многомерных векторных представлений слов [22-24] стимулирует появление большого объема работ по использованию многомерных векторных представлений слов для поиска на этапе извлечения [25-27]. В отличие от дискретного лексического представления, многомерное векторное представление слов являются непрерывным представлением, способным несколько облегчить проблему несоответствия словаря запроса и документа. После 2016 года наблюдается всплеск исследовательского интереса к применению методов глубокого машинного обучения для поиска на этапе извлечения [28-29]. Эти подходы изучены для улучшения представления документов в рамках традиционной парадигмы дискретного лексического представления [30-32] и непосредственно для формирования новых моделей семантического поиска в рамках парадигмы разреженного / плотного представления [33-36].

Отличия задач поиска по товарам от задач поиска по документам заключаются в следующем:

1) При извлечении информации о товарах иначе работают механизмы ранжирования на основании весов текстовых признаков (TF/IDF, BM25). Частота токена в названии товара не делает товар более релевантным запросу.
2) Товары обладают несколькими модальностями: название, описание, характеристики, изображение, видео. Поиск может производиться с учетом нескольких модальностей.

3) Поисковые запросы мотивированы интересом к товару и желанием его купить. Поведение покупателей отличается от поведения пользователя при поиске документа или интернет-ресурса.
4) Оценка эффективности поиска производится по модальностям.

Основной исследовательский вопрос – как влияет архитектура модели извлечения в части функции потерь и токенизации на автономные показатели системы извлечения товаров.

Далее исследовательский вопрос рассмотрен с точки зрения теории и практики, в заключении представлены основные выводы.

2. Методика

На рис. 1 изображено место системы извлечения товаров в системе поиска и отмечена граница взаимодействия с прямым и обратным индексами при онлайн-обработке поискового запроса.

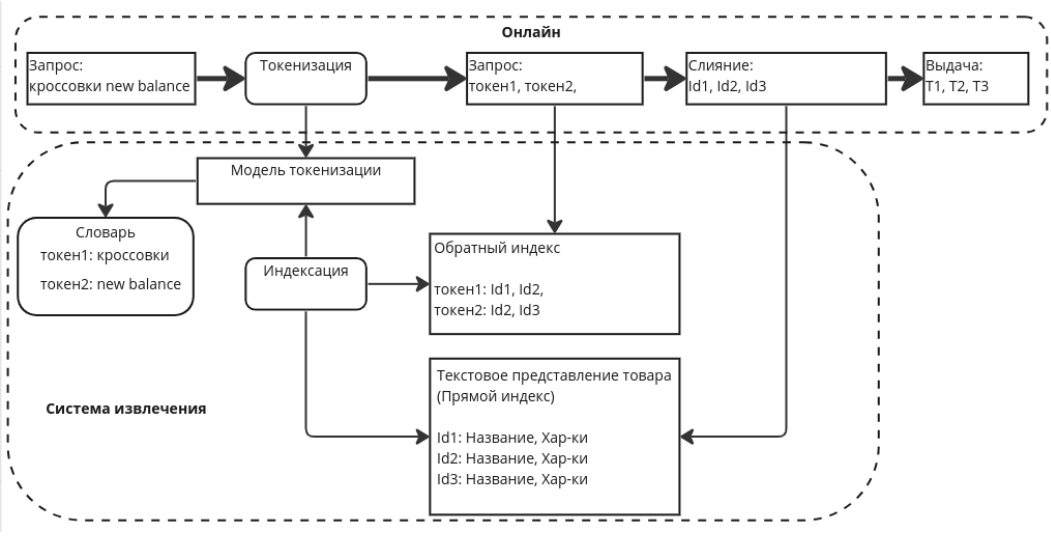

Рис. 1. Место системы извлечения информации о товарах в системе поиска.
Fig. 1. The place of the retireval system in the product search framework.

Развитие систем поиска товаров вместе с информационным поиском движется от лексических методов извлечения в сторону нейросетевых [37-39]. Одна из наиболее востребованных нейросетевых архитектур – DSSM [40]. Оригинальная архитектура DSSM получила развитие в виде парадигмы Dual Encoder [41-43] из-за двух независимых «башен», кодеров для поискового запроса и текстового представления товара. В Dual Encoder запросы и текстовое представление товара кодируются отдельно в общее многомерное векторное пространство фиксированного размера. Затем используется приближенный поиск [44-45] для извлечения релевантных товаров по многомерному векторному представлению поискового запроса. Наиболее перспективными стали исследования по применению моделей BERT в составе архитектуры Dual Encoder [46-48]. Общий принцип этих архитектур представлен в формулах 1–3.

$$\vec{q} = AvgPool\left[BERT_{\theta}^{l}(q)\right] \quad (1)$$

$$\vec{p} = AvgPool[BERT_{\theta}^{r}(p)] \quad (2)$$

$$s_{BERT}(\vec{q},\vec{p}) = \vec{q}^{T} \cdot \vec{p} \quad (3)$$

где $BERT_\theta^l$ и $BERT_\theta^r$ – это «левый» и «правый» кодеры, которые преобразуют тексты q и p в единое векторное пространство θ. Функция оценки соответствия $s_{BERT}(\cdot,\cdot)$ реализована через скалярное произведение $\vec{q}$ и $\vec{p}$. Узким местом данной архитектуры является усреднение векторов токенов по фрагменту текста.

Модель ColBERT [35] представляет частный случай Dual Encoder, который называют Single Encoder, так как используется один кодер и для запросов, и для товаров. Но основное новшество ColBERT состоит в том, что не используется усреднение по токенам AvgPool. Таким образом, на выходе BERT получаются векторы для каждого токена запроса и текстового представления продукта. Если q состоит из m токенов, а p состоит из n токенов, то функция оценки соответствия $s_{ColBERT}(\cdot,\cdot)$ будет следующая:

$$s_{ColBERT}(q_{1:m}, p_{1:n}) = \sum_1^m max_{1..n}\,(\vec{q}_{1:m}^T \cdot \vec{p}_{1:n}) \quad (4)$$

Суммирование в формуле (4) требует вычисления □ ∗ □ скалярных произведений по сравнению с одним произведением в $s_{BERT}(\cdot,\cdot)$.

Гибридизацию определяют как смешивание лексического и нейросетевого подходов к извлечению товаров, которое может быть произведено на разных уровнях. Например, как в исследовании [39], в выдаче смешаны результаты извлечения раздельными моделями, основанными на лексических, поведенческих и семантических методах. В исследовании [49] использован другой принцип создания гибридной модели: основной метод извлечения – лексический, и на ошибках лексической модели обучена семантическая модель.

Значительный рост автономных показателей моделей для работы с текстом достигнут за счет прогресса в методах токенизации [50-52]. Метод BPE (Byte-Pair-Encoding) изначально был представлен в литературе по сжатию данных [53]. BPE – это вариант словарного кодера, который постепенно находит набор символов таким образом, что общее количество символов для кодирования текста сводится к минимуму. Другая модель токенизации, называемая unigram, является энтропийным кодером, который минимизирует общую длину кода для текста. Согласно теореме Шеннона о кодировании, оптимальная длина кода для символа s равна – $log(p_s)$, где p_s – вероятность появления s. Фактически это то же самое, что стратегия сегментации языковой модели unigram, описанная в виде алгоритма Витерби [54]. BPE и unigram разделяют одну и ту же идею – кодирование текста с определенным принципом сжатия данных при использовании меньшего количества битов. Поэтому использование языковой модели unigram имеет те же преимущества, что и BPE. Для поиска по товарам имеют значение не токенизируемые на субсловарные единицы сущности, к примеру, бренды. Пользователь, ищущий товары бренда «Синий лён», не заинтересован в товарах содержащих токены «синий» и «лён» по отдельности.

Поясним разницу между задачами открытия новых брендов как сущностей и задачи использования существующих брендов для улучшения качества поиска. Задача открытия брендов объединяет несколько источников, таких как журналы пользовательских поисковых запросов, информацию о брендах из карточек товаров и реестры торговых марок для экспертной оценки является или нет последовательность слов брендом. Автоматизация процесс открытия брендов может быть выполнена с помощью моделей машинного обучения, но не является предметом данного исследования. В настоящем исследовании подразумевается, что мы имеем справочник брендов и ищем способы использовать его с максимальной пользой для поисковой системы.

С учетом вышеизложенного, в H1 реализован и опробован экспериментально гибридный подход, сочетающий разделения текста на словарные n-граммы различного размера без учета границ слов по пробелам. В качестве наиболее информативных токенов состоящих из нескольких слов, разделенных пробелами, для эксперимента взяты бренды, состоящие из нескольких слов (рис.2).

В H1 полученные векторы токенов не усредняются для получения близости, а по аналогии с ColBERT учитываются отдельно, как это показано на рис. 3.

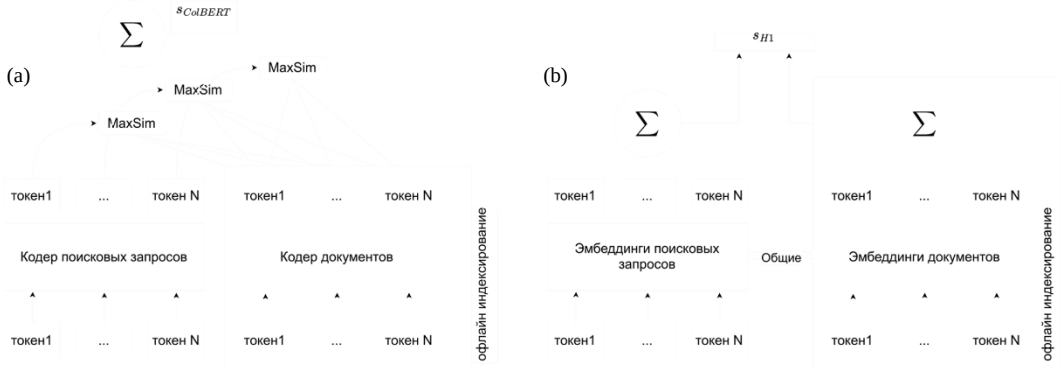

Рис. 2. (a). Общая схема работы ColBERT. (b). Общая схема работы H1.
Fig. 2. (a). ColBERT model. (b). H1 model.

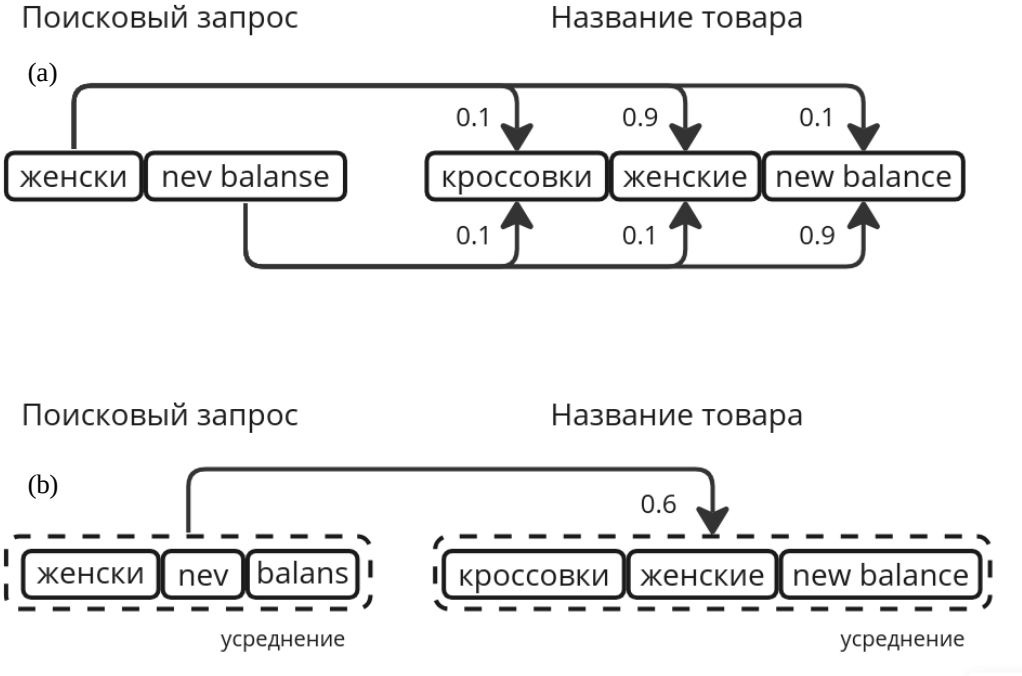

Рис. 3. (a). Оценка соответствия в H1 (b). Оценка соответствия в BERT.
Fig. 3. (a). Similarity in H1 (b). Similarity in BERT.

Методика оценки показателей эффективности для поиска по товарам отличается от оценки поиска по документам. Поиск по товарам подразумевает обнаружение нескольких, а желательно всех товаров, соответствующих запросу, даже если эти товары идентичны. Такое поведение продиктовано необходимостью сравнения цен на идентичные товары. Поэтому в формулах полноты и точности для сравнения полученной выдачи и истинной выдачи может использоваться совпадение как по текстовой модальности товара, так и по его цифровому идентификатору.

$$P_m@k = \frac{1}{|Q|}\sum_{q_i \in Q} \frac{\left|M_m\left(p^g_{q_i}, p^r_{q_i}@k\right)\right|}{k} \quad (5)$$

$$mAP_m@k = \frac{1}{k}\sum_{i=1}^{k} P_m@i \quad (6)$$

$|Q|$ – количество рассматриваемых поисковых запросов,

@k, @i – порог отсечки поисковой выдачи,

q_i – поисковый запрос $q_i \in Q$,

$p_{q_i}^{g}$ – все товары, соответствующие поисковому запросу q_i,

$p_{q_i}^{r}$ – поисковая выдача, все товары, найденные по поисковому запросу q_i

$|M_m(\cdot,\cdot)|$ – количество соответствующих друг другу товаров, определенных по функции соответствия M для модальности m.

Для более глубокого исследования возможностей оценки эффективности систем продуктового поиска по показателям можно ознакомиться с исследованием автора [56].

2. Эксперимент

Для подтверждения обозначенных преимуществ модели извлечения H1 выбран публичный набор данных, обучено три модели токенизации и проведены сравнительные эксперименты с моделью извлечения H1 и моделями TCT-ColBERT[55], Single Encoder (SE) [39] и Dual Encoder (DE) [40] на основе показателей mAP@12 и R@1k (5).

В качестве набора данных выбран WANDS [57], позволяющий проводить объективный бенчмаркинг и оценку систем извлечения товаров на основе набора данных электронной коммерции. Его ключевые характеристики включают:

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

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

Для сравнения выбрана реализация модели ColBERT от Terrier [58], обученная с использованием подхода Tight Coupling Teachers [55].

В качестве гиперпараметров для моделей H1, SE, DE взяты значения размерности векторов токенов: 128, 256, 512, 768. В качестве функции соответствия M использовано точное соответствие для модальности m. Модальность m определена как «полное название товара».

Этапы проведения эксперимента:

1) Построение моделей токенизации word, bpe, unigram с токенами из нескольких слов (mt) и только с субсловарными конструкциями (non mt);
2) Обучение моделей извлечения товаров с различными вариантами векторного представления токенов;
3) Индексирование товаров с помощью модели извлечения товаров;
4) Вычисление показателей точности и полноты для каждого поискового запроса с различными порогами от 1 до 4096.

В общей сложности обработано более 300 тысяч вариантов комбинаций. Результаты экспериментов, усредненные по поисковым запросам, представлены на рис 4–9.

Наилучшие показатели R@1k = 86.6% и mAP@12 = 56.1% получены для модели H1 с моделью токенизации BPE, размерностью векторов 768 и токенами из «нескольких слов» (mt). Для сравнения на тех же данных модель ColBERT показала значения R@1k = 84.0% и

mAP@12 = 42.1%. Показатели пороговой полноты и точности для модели извлечения товаров на основе ColBERT приведены в табл. 1.

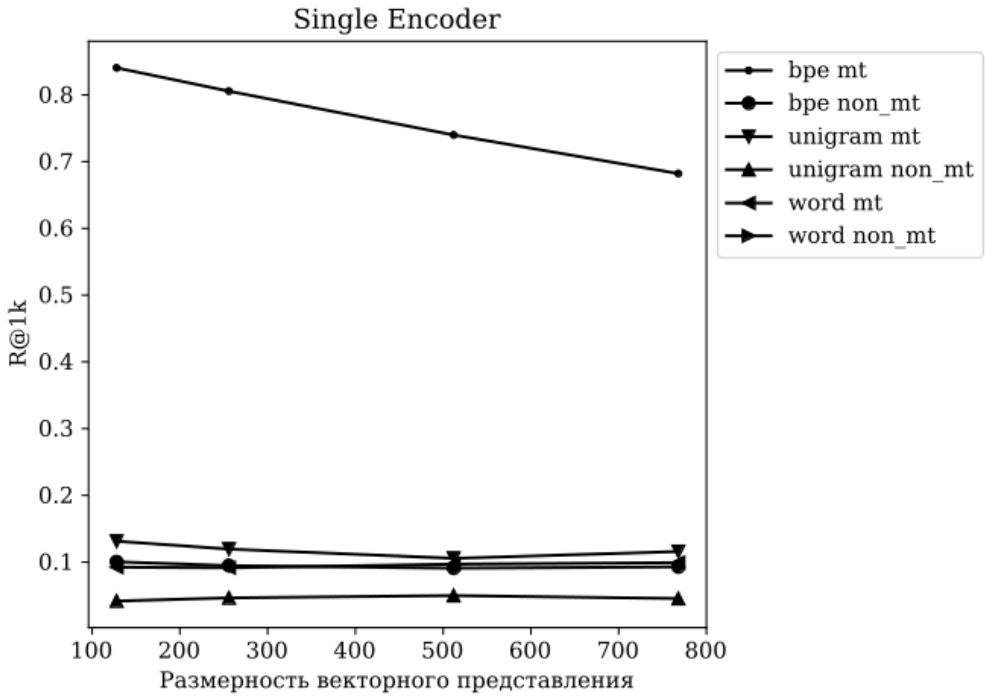

Рис. 4. Зависимость показателя R@1k от гиперпараметров для модели Single Encoder.
Fig. 4. Dependence of the R@1k metric on hyperparameters for the Single Encoder model.

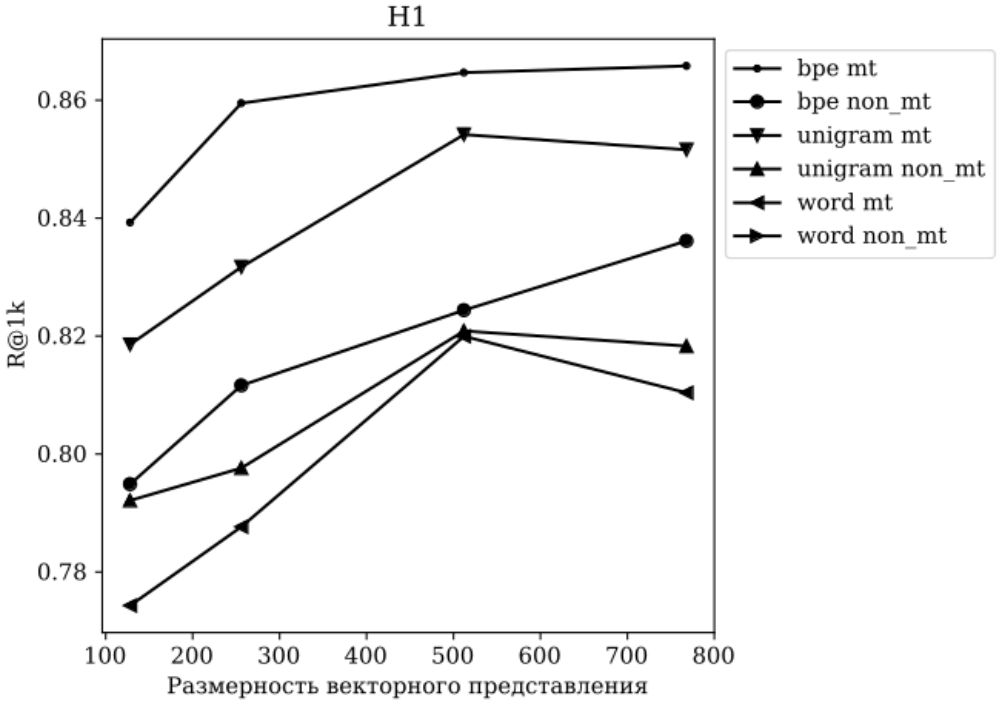

Рис. 5. Зависимость показателя R@1k от гиперпараметров для модели H1.
Fig. 5. Dependence of the R@1k metric on hyper parameters for the H1 model.

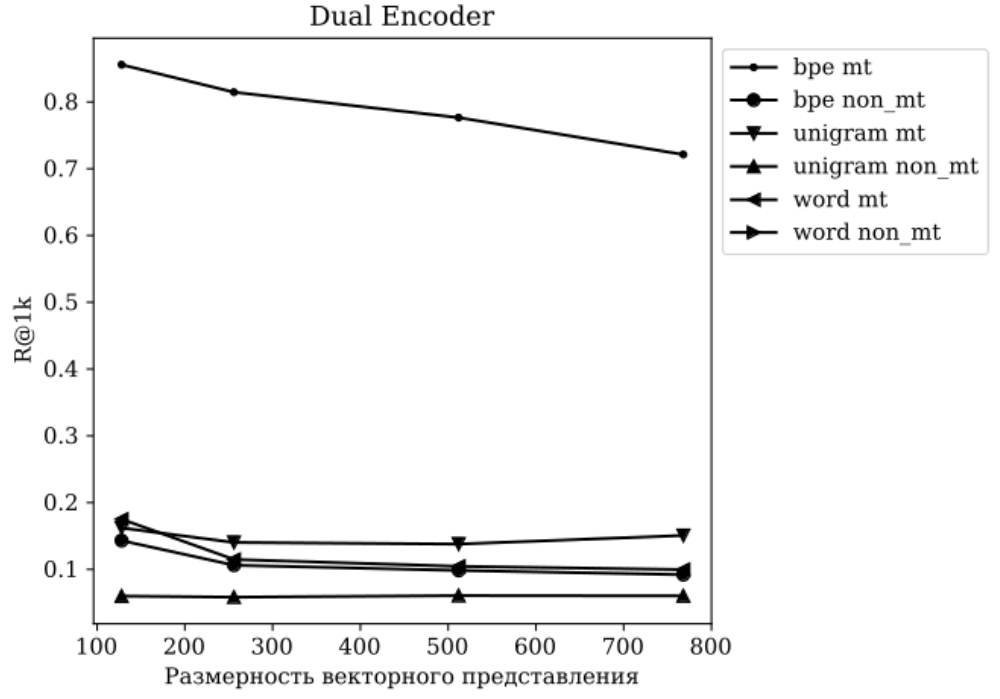

Рис. 6. Зависимость показателя R@1k от гиперпараметров для модели Dual Encoder.
Fig. 6. Dependence of the R@1k metric on hyperparameters for the Dual Encoder model.

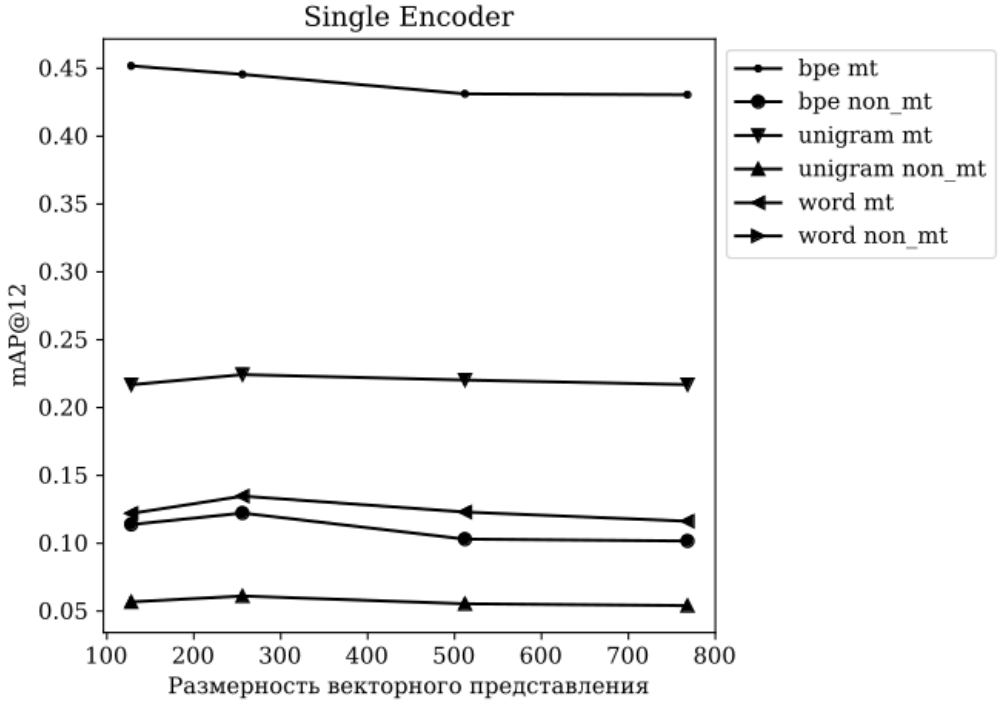

Рис. 7. Зависимость показателя mAP@12 от гиперпараметров для модели Single Encoder.
Fig. 7. Dependence of the mAP@12 metric on hyperparameters for the Single Encoder model.

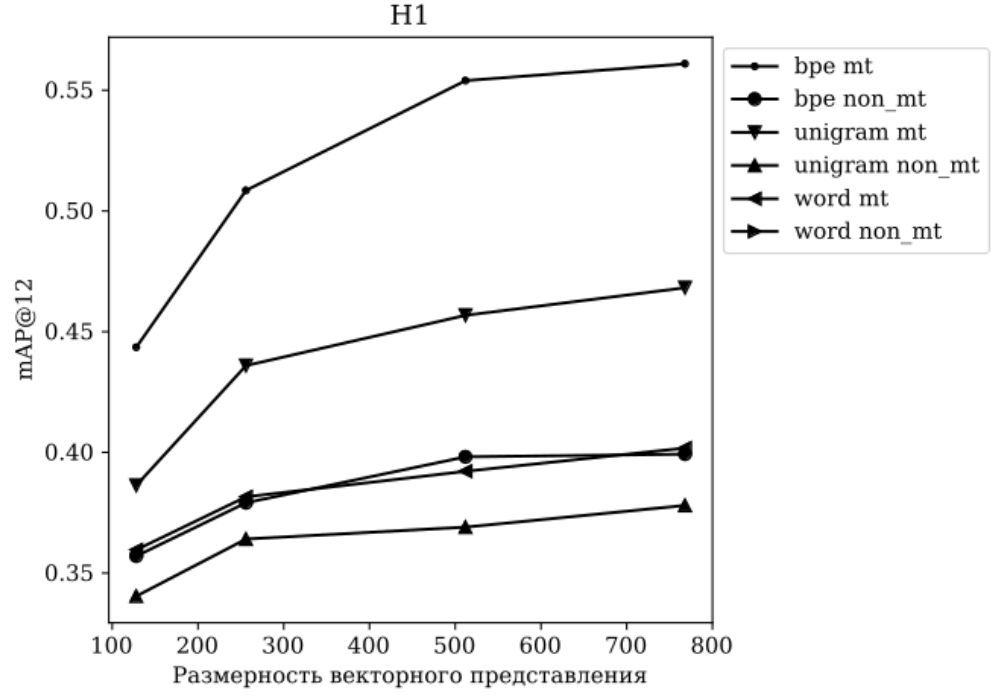

Рис. 8. Зависимость показателя mAP@12 от гиперпараметров для модели H1.
Fig. 8. Dependence of the mAP@12 metric on hyperparameters for the H1 model.

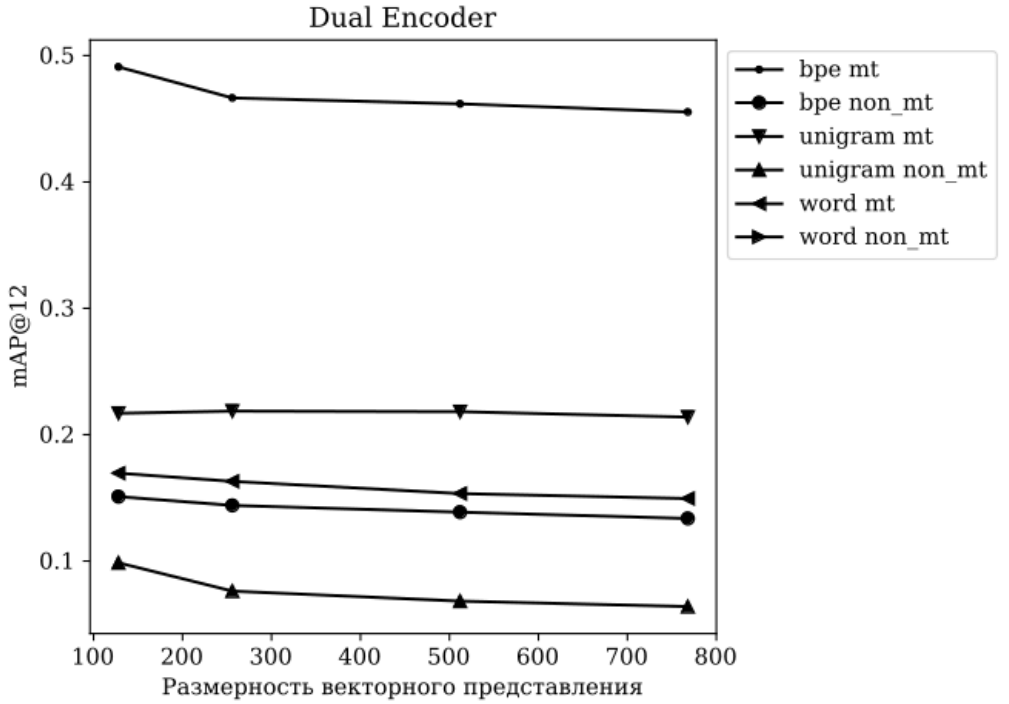

Рис. 9. Зависимость показателя mAP@12 от гиперпараметров для модели Dual Encoder.
Fig. 9. Dependence of the mAP@12 metric on hyperparameters for the Dual Encoder model.

Табл. 1. Показатели пороговой полноты и точности для модели ColBERT.
Table 1. R@k and P@k for the ColBERT model.

Модель	Порог, k	Точность	Полнота
ColBERT	12	**41%**	26%
	128	21%	61%
	512	9%	78%
	1024	5%	**84%**

Для получения результатов, приведенных на рис. 3–8, была использована разметка примеров из набора данных и функция потерь на основе кросс-энтропии. В промышленных масштабах разметить достаточное количество данных не представляется возможным из-за их высокой скорости изменения, поэтому модели извлечения товаров, обучаемые на поведенческих данных, используют правила для генерации отрицательных примеров.

Рис 3–8 сделаны с усреднением показателей mAP@12 и R@1k по запросам, чтобы сравнить модели и гиперпараметры. Для лучшего набора гиперпараметров на рис. 9 приведены зависимости показателей полноты и точности для рассматриваемых моделей без усреднения для одного запроса.

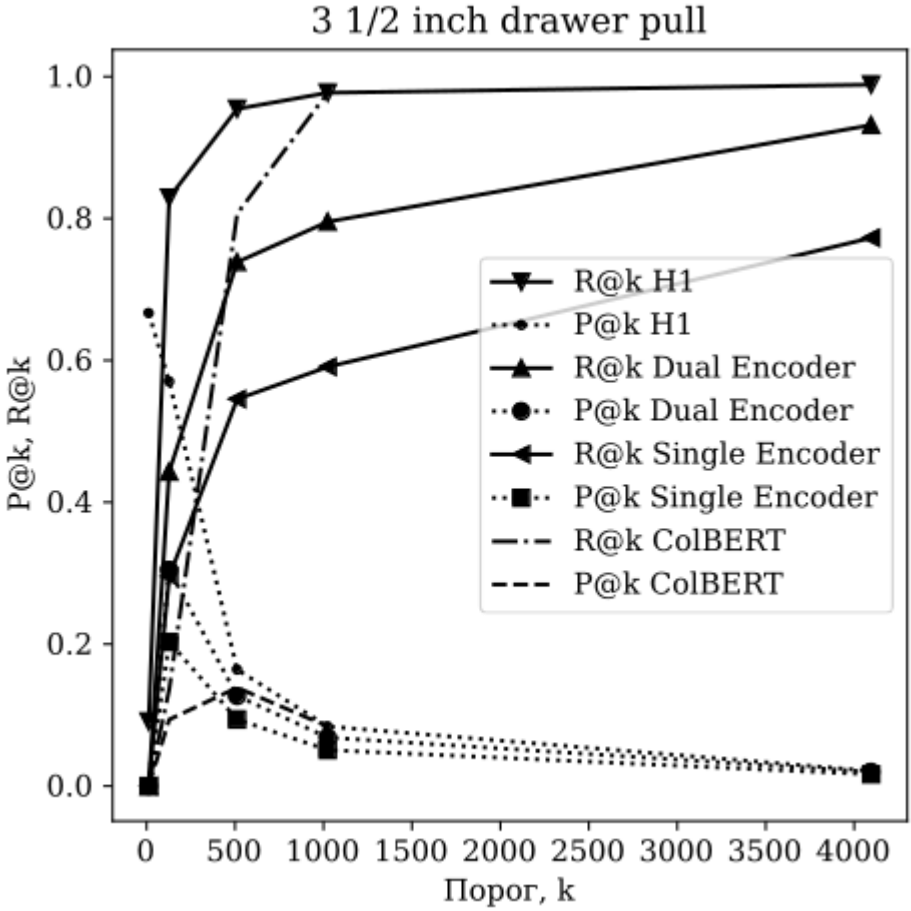

Рис. 10. Зависимости показателей полноты и точности от порога для одного поискового запроса.
Fig. 10. Dependences of recall and precision metrics for a single search query.

Зависимости для показателя полноты и точности на рисунке 10 демонстрируют преимущества модели извлечения товаров H1 над моделями Single Encoder, ColBERT и Dual Encoder. Выдача становится более полной при меньших значениях порога k, а точность выдачи при этом снижается медленнее.

4. Заключение

Классические модели лексического поиска балансируют между крайне общими и слишком специфичными поисковыми запросами, как следствие, необъятной или пустой поисковой выдачей. Подобный недостаток отдаляет клиентов от покупки товара на электронной торговой интернет-площадке. Модели поиска, основанные на векторном представлении, способны мягко сопоставлять запросы и документы, но они теряют информацию о соотнесении на уровне конкретных слов. В данной статье представлена модель поиска H1, которая лексический поиск дополняет поиском с многомерным векторным пространствам. Экспериментально доказано, что H1 достигает высокой эффективности поиска на первом этапе в рамках публичного набора данных, превосходя подходы на основе BERT за счет легкой архитектуры и уменьшения размера словаря с помощью употребления токенов, состоящих из нескольких слов. При этом H1 способна генерировать детализированные

сигналы сопоставления на уровне токенов, которые имеют решающее значение для точного поиска.

Список литературы / References

[1]. Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. pp. 232–241 (1994)

[2]. Zeng C. et al. FAERY: An FPGA-accelerated Embedding-based Retrieval System //16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22). – 2022. – С. 841-856.

[3]. Zeng S. et al. DF-GAS: a Distributed FPGA-as-a-Service Architecture towards Billion-Scale Graph-based Approximate Nearest Neighbor Search //Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture. – 2023. – С. 283-296.

[4]. Pan Z. et al. RECom: A Compiler Approach to Accelerate Recommendation Model Inference with Massive Embedding Columns. – 2023.

[5]. Hofstätter S. et al. Improving efficient neural ranking models with cross-architecture knowledge distillation //arXiv preprint arXiv:2010.02666. – 2020.

[6]. George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. 1987. The vocabulary problem in human-system communication. Commun. ACM 30, 11 (1987), 964–971.

[7]. Le Zhao and Jamie Callan. 2010. Term necessity prediction. In Proceedings of the 19th ACM international conference on Information and knowledge management. 259–268.

[8]. Hang Li and Jun Xu. 2014. Semantic matching in search. Foundations and Trends in Information retrieval 7, 5 (2014), 343–469.

[9]. Victor Lavrenko and W. Bruce Croft. 2001. Relevance Based Language Models. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (New Orleans, Louisiana, USA) (SIGIR '01). Association or Computing Machinery, New York, NY, USA, 120–127. https://doi.org/10.1145383952.383972

[10]. Michael E Lesk. 1969. Word-word associations in document retrieval systems. American documentation 20, 1 (1969), 27–38.

[11]. Yonggang Qiu and Hans-Peter Frei. 1993. Concept Based Query Expansion. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Pittsburgh, Pennsylvania,USA) (SIGIR '93). Association for Computing Machinery, New York, NY, USA, 160–169. https://doi.org/10.1145/160688.160713

[12]. Jinxi Xu and W Bruce Croft. 2017. Quary expansion using local and global document analysis. In Acm sigir forum,Vol. 51. ACM New York, NY, USA, 168–175.

[13]. Miles Efron, Peter Organisciak, and Katrina Fenlon. 2012. Improving retrieval of short texts through document expansion. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. 911–920.

[14]. Xiaoyong Liu and W Bruce Croft. 2004. Cluster-based retrieval using language models. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. 186–193.

[15]. Jianfeng Gao, Jian-Yun Nie, Guangyuan Wu, and Guihong Cao. 2004. Dependence Language Model for Information Retrieval. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Sheffield, United Kingdom) (SIGIR '04). Association for Computing Machinery, New York, NY,USA, 170–177. https://doi.org/10.1145/1008992.1009024

[16]. Donald Metzler and W. Bruce Croft. 2005. A Markov Random Field Model for Term Dependencies. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Salvador, Brazil) (SIGIR '05). Association for Computing Machinery, New York, NY, USA, 472–479. https://doi.org/10.1145/1076034.1076115

[17]. Jun Xu, Hang Li, and Chaoliang Zhong. 2010. Relevance ranking using kernels. In Asia Information Retrieval Symposium. Springer, 1–12.

[18]. Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. Journal of the American society for information science 41, 6 (1990), 391–407.

[19]. Xing Wei and W. Bruce Croft. 2006. LDA-Based Document Models for Ad-Hoc Retrieval. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information

Retrieval (Seattle, Washington,USA) (SIGIR '06). Association for Computing Machinery, New York, NY, USA, 178–185. https://doi.org/10.1145/1148170.1148204

[20]. Adam Berger and John Lafferty. 1999. Information Retrieval as Statistical Translation. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Berkeley, California,USA) (SIGIR '99). Association for Computing Machinery, New York, NY, USA, 222–229. https://doi.org/10.1145/312624.312681

[21]. Maryam Karimzadehgan and ChengXiang Zhai. 2010. Estimation of Statistical Translation Models Based on Mutual Information for Ad Hoc Information Retrieval. In Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (Geneva, Switzerland) (SIGIR '10). Association for Computing Machinery, New York, NY, USA, 323–330. https://doi.org/10.1145/1835449.1835505

[22]. Felipe Bravo-Marquez, Gaston L'Huillier, Sebastián A. Ríos, and Juan D. Velásquez. 2010. Hypergeometric Language Model and Zipf-like Scoring Function for Web Document Similarity Retrieval. In Proceedings of the 17th International Conference on String Processing and Information Retrieval (Los Cabos, Mexico) (SPIRE'10). Springer-Verlag, Berlin,Heidelberg, 303–308.

[23]. Mikolov T. et al. Distributed representations of words and phrases and their compositionality //Advances in neural information processing systems. – 2013. – T. 26.

[24]. Pennington J., Socher R., Manning C. D. Glove: Global vectors for word representation //Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). – 2014. – C. 1532-1543.

[25]. Clinchant S., Perronnin F. Aggregating continuous word embeddings for information retrieval //Proceedings of the workshop on continuous vector space models and their compositionality. – 2013. – C. 100-109.

[26]. Ganguly D. et al. Word embedding based generalized language model for information retrieval //Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval. – 2015. – C. 795-798.

[27]. Vulić I., Moens M. F. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings //Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval. – 2015. – C. 363-372.

[28]. Boytsov L. et al. Off the beaten path: Let's replace term-based retrieval with k-nn search //Proceedings of the 25th ACM international on conference on information and knowledge management. – 2016. – C. 1099-1108.

[29]. Henderson M. et al. Efficient natural language response suggestion for smart reply //arXiv preprint arXiv:1705.00652. – 2017.

[30]. Bai Y. et al. SparTerm: Learning term-based sparse representation for fast text retrieval //arXiv preprint arXiv:2010.00768. – 2020.

[31]. Dai Z., Callan J. Context-aware sentence/passage term importance estimation for first stage retrieval //arXiv preprint arXiv:1910.10687. – 2019.

[32]. Nogueira R. et al. Document expansion by query prediction //arXiv preprint arXiv:1904.08375. – 2019.

[33]. Gillick D., Presta A., Tomar G. S. End-to-end retrieval in continuous space //arXiv preprint arXiv:1811.08008. – 2018.

[34]. Jean S. et al. On using very large target vocabulary for neural machine translation //arXiv preprint arXiv:1412.2007. – 2014.

[35]. Khattab O., Zaharia M. Colbert: Efficient and effective passage search via contextualized late interaction over bert //Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval. – 2020. – C. 39-48.

[36]. Zamani H. et al. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing //Proceedings of the 27th ACM international conference on information and knowledge management. – 2018. – C. 497-506.

[37]. Li S. Embedding-based product retrieval in Taobao search // Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. – 2021. – C. 3181-3189.

[38]. Magnani A. Semantic retrieval at Walmart // Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. – 2022. – C. 3495-3503.

[39]. Nigam P. et al. Semantic product search //Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. – 2019. – C. 2876-2885.

[40]. Huang P. S. et al. Learning deep structured semantic models for web search using clickthrough data //Proceedings of the 22nd ACM international conference on Information & Knowledge Management. – 2013. – С. 2333-2338.
[41]. Gillick D., Presta A., Tomar G. S. End-to-end retrieval in continuous space //arXiv preprint arXiv:1811.08008. – 2018.
[42]. Yang Y. et al. Multilingual universal sentence encoder for semantic retrieval //arXiv preprint arXiv:1907.04307. – 2019.
[43]. Karpukhin V. et al. Dense Passage Retrieval for Open-Domain Question Answering //EMNLP (1). – 2020. – С. 6769-6781.
[44]. Vanderkam D. et al. Nearest neighbor search in google correlate. – 2013.
[45]. Johnson J., Douze M., Jégou H. Billion-scale similarity search with gpus //IEEE Transactions on Big Data. – 2019. – Т. 7. – №. 3. – С. 535-547.
[46]. Chang W. C. et al. Pre-training tasks for embedding-based large-scale retrieval //arXiv preprint arXiv:2002.03932. – 2020.
[47]. Xiong L. et al. Approximate nearest neighbor negative contrastive learning for dense text retrieval //arXiv preprint arXiv:2007.00808. – 2020.
[48]. Lu W., Jiao J., Zhang R. Twinbert: Distilling knowledge to twin-structured compressed bert models for large-scale retrieval //Proceedings of the 29th ACM International Conference on Information & Knowledge Management. – 2020. – С. 2645-2652.
[49]. Gao L. et al. Complement lexical retrieval model with semantic residual embeddings //Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part I 43. – Springer International Publishing, 2021. – С. 146-160.
[50]. Kudo T. Subword regularization: Improving neural network translation models with multiple subword candidates //arXiv preprint arXiv:1804.10959. – 2018.
[51]. Sennrich R., Haddow B., Birch A. Neural machine translation of rare words with subword units //arXiv preprint arXiv:1508.07909. – 2015.
[52]. Provilkov I., Emelianenko D., Voita E. BPE-dropout: Simple and effective subword regularization //arXiv preprint arXiv:1910.13267. – 2019.
[53]. Gage P. A new algorithm for data compression //C Users Journal. – 1994. – Т. 12. – №. 2. – С. 23-38.
[54]. Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm //IEEE transactions on Information Theory. – 1967. – Т. 13. – №. 2. – С. 260-269.
[55]. Lin S. C., Yang J. H., Lin J. Distilling dense representations for ranking using tightly-coupled teachers //arXiv preprint arXiv:2010.11386. – 2020.
[56]. Krasnov F.V. Embedding-based retrieval: measures of threshold recall and precision to evaluate product search. // Business Informatics. – 2024 – Т. 18. – №. 2. – С.22–34. DOI:10.17323/2587-814X.2024.2.22.34.
[57]. Chen Y. Wands: Dataset for product search relevance assessment // European Conference on Information Retrieval. – Cham : Springer International Publishing, 2022. – С. 128-141.
[58]. Macdonald C., Tonellotto N. Declarative experimentation in information retrieval using PyTerrier //Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval. – 2020. – С. 161-168.

Информация об авторах / Information about authors

Федор Владимирович КРАСНОВ – доктор технических наук, специалист по поисковым и рекомендательным системам в электронной коммерции, сотрудник Исследовательского центра ООО "ВБ СК" на базе Инновационного Центра Сколково.

Fedor Vladimirovich KRASNOV – Dr. 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, Moscow