Исследование влияния текстового представления товара посредством моделей с использованием искусственных нейронных сетей глубокого обучения на релевантность поиска товаров на электронной торговой Интернет-площадке

Ф.В.Краснов

***Аннотация*— С ростом популярности онлайн-покупок академические исследования в области электронной коммерции обретают все большую актуальность. Тем не менее, сохраняются значительные исследовательские проблемы, начиная с классических проблем поиска в электронной коммерции, например, сопоставление текстовых запросов с мультимодальными представлениями товаров, оптимизация ранжирования для двусторонних торговых Интернет-площадок с учетом взаимодействия покупателя и продавца с рекомендательными и поисковыми системами. Эти области исследований важны для понимания поведения клиентов, стимулирования вовлеченности и повышения конверсий. Данные о товаре играют ключевую роль в поиске и рекомендациях. Сервисы, в которых на электронной торговой Интернет-площадке есть несколько поставщиков (маркетплейсы), демонстрируют высокую динамику в отношении качества и согласованности контента, выявления мошенничества и ценообразования. Обновления запасов товаров часто сопровождаются соглашениями об уровне обслуживания, гарантирующими внесение изменений в соответствии с требованиями клиентов в строго установленные сроки. Наконец, сами товары могут рассматриваться как мультимодальные документы, и успех поиска зависит от соответствия намерений клиента всем аспектам товара. Для конкретности мы будем использовать термин «документы» для элементов, возвращаемых по поисковому запросу, хотя возвращаемые элементы могут быть более общими (например, мультимедийные элементы).**

При поиске на электронной торговой Интернет-площадке индексированные товары, которые ищут покупатели, представляют собой комбинации изображений, видео и неструктурированного текста (названия, описания и обзоры) и структурированных данных (цена, бренд, рейтинги, местоположение продавца, логистика). Это сочетание данных открывает перспективы для исследований, включая улучшение извлечения данных из разнообразных источников с использованием сигналов из различных типов данных, таких как предоставление более подробной информации о цвете и использование сходства изображений для рекомендаций; а также как способ для клиентов создавать запросы в поисковой системе. Данные являются ключом к обеспечению высококачественного поиска и рекомендаций, отвечающих потребностям клиентов и бизнеса. В данной статье автор исследовал возможности нескольких подходов к извлечению полезных сигналов из различных источников информации о товаре для улучшения клиентского опыта в сфере электронной коммерции.

***Ключевые слова* — NDCG, DNN, transformers, BERT, e-commerce, Information Retrieval, IR.**

I. Введение

Приложения для поиска и рекомендаций варьируются от традиционного веб-поиска по коллекции текстовых документов до вертикальных многоуровневых поисковых систем [1]. В статье автор исследует подходы к поиску товаров в сфере электронной коммерции. Несмотря на то, что основная задача поиска (т.е. удовлетворение информационных потребностей пользователя) такая же, как и при веб-поиске, способ ее решения отличается. На сайтах электронной коммерции данные, доступные для поиска и ранжирования, разнятся, как и сигналы успеха (например, добавление товаров в корзину, покупка). Объекты, которые необходимо обнаружить (товары), представляют собой комбинации неструктурированного текста (заголовки, описания, обзоры), изображений и структурированных данных (цена, бренд, рейтинги, популярность, доход). Это сложное сочетание данных ставит интересные исследовательские задачи, включая функции сопоставления и ранжирования, которые учитывают компромиссы между аспектами в отношении запроса пользователя. Функции, доступные для построения моделей кликов, используемых при ранжировании, в электронной коммерции отличаются и часто являются более эффективными, чем при веб-поиске [2]. Помимо запросов, времени наведения курсора мыши, кликов и времени просмотра, сайты электронной коммерции получают сигналы о добавлении в корзину, покупке, параллельном сравнении, удалении из корзины, возврате товаров и т.д. При включении рекламных акций и персонализации, к примеру, индивидуального

ценообразования, модели кликов становятся более сложными [3], чем при веб-поиске. Кроме того, электронная коммерция характеризуется динамичным ассортиментом с высокой скоростью изменений и оборачиваемости, а также очень длинным «хвостом» распределения частот поисковых запросов [4].

II. Методика

Ранжирование широко изучено в области информационного поиска, машинного обучения и статистики, так как оно играет центральную роль в поисковых, рекомендательных и экспертных системах. Фундаментальная проблема заключается в том, как разработать показатель ранжирования для оценки эффективности функции ранжирования. В отличие от классификации и регрессии, для которых существуют простые и естественные метрики, оценка функций ранжирования является более сложной задачей. Предположим, что существует n объектов для ранжирования. Мера оценки ранжирования должна приводить к общему упорядочению n возможных результатов ранжирования. Однако существует много способов определения показателей ранжирования, и было предложено несколько мер оценки [5,6,7,8]: ожидаемый взаимный ранг (expected reciprocal rank, ERR) [9] и взвешенный прирост информации (weighted information gain, WIG) [10].

На самом деле, как отмечают авторы исследования [11], не существует единого оптимального показателя ранжирования, который работал бы для любого приложения.

В центре внимания данной работы находится нормализованный дисконтированный совокупный выигрыш (NDCG), который представляет собой семейство показателей ранжирования, широко используемых в приложениях [12]. NDCG имеет два преимущества по сравнению со многими другими мерами. Во-первых, NDCG позволяет каждому извлеченному документу иметь ранжированную релевантность, в то время как большинство традиционных мер ранжирования допускают только бинарную релевантность. То есть каждый документ рассматривается либо как релевантный, либо как нерелевантный по предыдущим показателям ранжирования, тем временем для документов в NDCG доступны разные степени релевантности. Во-вторых, NDCG включает функцию дисконтирования по рангу, при этом многие другие показатели равномерно взвешивают все позиции. Эта функция особенно важна для поисковых систем, так как пользователям наиболее интересны документы с самым высоким рейтингом.

NDCG – это нормализованная форма показателя дисконтированного совокупного выигрыша (DCG). DCG – это взвешенная сумма степени релевантности ранжированных элементов. Вес является убывающей функцией ранга (положения) объекта и поэтому называется дисконтированием. Первоначальная причина для введения дисконтирования – уменьшение вероятности просмотра пользователем документа в зависимости от его ранга. NDCG нормализует DCG по идеальному DCG (IDCG), который является показателем наилучшего результата ранжирования. Таким образом, мера NDCG всегда является числом в промежутке [0, 1]. Строго говоря, NDCG – это семейство показателей ранжирования, поскольку существует определенная свобода при выборе функции выигрыша. В научной литературе и программных библиотеках преобладает логарифмический метод дисконтирования $\frac{1}{log(1+r)}$, где r – это ранг. Далее будет использован именно этот вариант как стандартный NDCG. Другая функция дисконтирования, появившаяся в научной литературе, это r^{-1}, которая называется Zipfian при поиске информации [13]. Поисковые системы также используют сокращенную версию NDCG top-k. В ней дисконт устанавливается равным нулю для рангов r, превышающих k. Такая мера NDCG обычно называется NDCG@k. Несмотря на широкое распространение метрики NDCG, у нее имеется ряд ограничений:

- Нормализованная метрика DCG не наказывает за плохие документы в результате. Например, если запрос возвращает два результата с оценками 1,1,1 и 1,1,1,0 соответственно, оба будут считаться одинаково хорошими, даже если последний содержит плохой документ. Для ранжирования оценок «Отлично», «Справедливо», «Плохо» можно использовать числовые оценки 1,0,-1 вместо 2,1,0. Это вызовет снижение оценки в случае возврата плохих результатов, отдавая приоритет точности результатов над отзывом. Обратим внимание, что такой подход может привести к общему отрицательному баллу, понижающему нижнюю границу балла с 0 до отрицательного значения.
- Нормализованная метрика DCG не наказывает за отсутствие документов в результате. К примеру, если запрос возвращает два результата с оценками 1,1,1 и 1,1,1,1,1 соответственно, оба будут считаться в равной степени хорошими, предполагая, что идеальный DCG вычисляется с рангом 3 для первого и рангом 5 для второго. Один из способов учесть это ограничение – установить фиксированный размер результирующего набора и использовать минимальные баллы для отсутствующих документов. В предыдущем примере автор использовал бы оценки 1,1,1,0,0 и 1,1,1,1,1 и указал NDCG как NDCG@5.
- Нормализованная метрика DCG может не подходить для измерения производительности запросов, которые часто способны давать несколько одинаково хороших результатов. Что особенно верно, когда данный показатель ограничен исключительно первыми несколькими результатами, как это делается на практике. Например, для таких запросов, как «юбка», NDCG@1 будет учитывать только первый результат, и, следовательно, если один

результирующий набор содержит всего одну “юбку” из близлежащего района доставки, в то время как другой содержит 5, оба в конечном итоге получат одинаковую оценку, даже несмотря на то, что последний является более полным.

Один из самых сложных вопросов в электронной коммерции – как ранжировать результаты поиска, показываемые клиентам. Ранжирование в электронной коммерции началось с сортировки товаров по названию или цене, а позже было расширено за счет базового логического поиска. Ученые и практики приложили большие усилия для получения единой функции ранжирования, которая смешивает логические алгоритмы ранжирования или алгоритмы ранжирования на основе текстовых признаков, к примеру TF-IDF, с другими сигналами, такими как новизна или популярность. Это дало неоднозначные результаты, потому что некоторые сигналы работают для одних запросов, и не работают для других. Например, по причинам, связанным с распределением товаров в каталоге, полосатые футболки по запросу могут ранжировать товары с полосой, отличные от футболок, потому что «полосатый» имеет значительно более высокий IDF, чем «футболки». По мере увеличения количества сигналов становится невозможным вывести функцию ранжирования, которая возвращает оптимальные результаты по всему спектру запросов, намерений клиентов и бизнес-целей. В последние годы метод «Обучение ранжированию» (Learning to Rank) все чаще используется для решения такого рода сложных задач оптимизации либо поточечным способом (путем минимизации потерь по сравнению с золотым стандартом), либо попарным способом (путем минимизации потерь, вызванных превышением допустимых значений пары элементов), либо же списковым способом (путем минимизации списочной функции потерь, определенной в прогнозируемом и истинном списках).

Запросы и документы могут быть представлены лексически и семантически. Эти представления должны совместно использоваться запросами и документами, так как они являются основой сопоставления. Помимо точности соответствия условиям запроса, документы имеют много иных характеристик, а именно, сколько раз они были куплены, сколько раз на них нажимали («клики»), как соотносятся клики и покупки. Системы способны расширять семантическое представление документа за счет включения существующих или создания новых характеристик. Семантическое пространство представления можно дополнительно расширить для захвата отношений более высокого порядка, например, ранга документа в категории в сочетании с оценкой сходства запроса и документа с одной или несколькими функциями сходства. Создание широкого и богатого пространства представления документов позволяет системам Learning to Rank изучать более сложные определения целевых функций.

Поисковые и рекомендательные системы электронной коммерции необходимо одновременно оптимизировать по нескольким критериям (также известным как целевые функции) [14]: один кодирует предпочтения клиентов, а другой – бизнес-предпочтения. Эти целевые функции конкурируют друг с другом, поднимая исследовательские вопросы относительно наилучшей стратегии обучения в краткосрочной и долгосрочной перспективе. Прежде чем разрабатывать функции оптимизации, отражающие удовлетворенность клиентов и успех в бизнесе, изучают сигналы, которые фиксируют цели каждого заинтересованного лица в отдельности. Удовлетворенность клиентов измеряется различными сигналами [15], включая рейтинг кликов, время наведения и задержки, удовлетворенные клики, переформулировку запроса, продолжительность сеанса, количество запросов до оформления заказа, добавление в корзину, покупки, время до следующего визита, возврат товара и звонков в службу поддержки клиентов. Успех бизнеса, в свою очередь, измеряется несколькими ключевыми показателями эффективности (KPI), включая:

- показатели, ориентированные на запасы (такие как среднее время оборота, количество уникальных товаров, проданных за период времени),
- показатели, ориентированные на доход,
- показатели, ориентированные на получение прибыли,
- показатели посещаемости (например, общее и уникальное количество посетителей, количество новых и вернувшихся посетителей),
- показатели, ориентированные на корзину (а именно: среднее количество товаров в корзине, средняя стоимость корзины).

Каждый сигнал может быть преобразован в целевую функцию, а их линейная интерполяция дает возможность определить глобальную целевую функцию. Учитывая свои краткосрочные и долгосрочные цели, а также маркетинговые решения, предприятия способны взвесить отдельные целевые функции и изучить метацелевую функцию, которая оптимизирует комбинацию целевых.

Метрику NDCG можно определить как меру эффективности системы или как меру удовлетворенности пользователя работой системы. Иногда такое деление метрик называют off-line и on-line. Для оценки удовлетворенности пользователя работой системы (on-line) используют бинарные и многоуровневые градации оценок товаров в выдачи по степени связанности с поисковым запросом.

В работе [16] приведена следующая шкала оценок релевантности:

- Точный (Т): товар релевантен запросу и удовлетворяет всем спецификациям запроса (например, бутылка для воды, соответствующая всем атрибутам запроса «пластиковая бутылка для воды 1 литр», таким как материал и размер).
- Заменитель (З): элемент в некоторой степени релевантен, т.е. он не удовлетворяет некоторым

аспектам запроса, но элемент может быть использован в качестве функциональной замены (к примеру, флис для запроса «свитер»).
- Дополнение (Д): товар не соответствует запросу, но может быть использован в сочетании с конкретным товаром (например, спортивные брюки для запроса «кроссовки»).
- Нерелевантный (I): товар не имеет отношения к делу или не соответствует центральному аспекту запроса (допустим, носки для запроса «телескоп» или хлеб из пшеничной муки для запроса «хлеб без глютена»).

Целевые функции по нескольким сигналам могут смещаться в сторону более частых сигналов. По своей природе одни сигналы более распространены, чем другие, и некоторые из них являются более сильными индикаторами предпочтения. Например, покупка является более явным индикатором предпочтений, чем клик, но встречается гораздо реже. Покупка, которую не вернули, является более сильным сигналом, чем покупка, но опять же встречается реже. Чем сильнее индикатор предпочтений, тем ниже амплитуда сигнала. Целевые функции должны учитывать эту разницу в силе сигнала по сравнению с количеством сигналов, в противном случае более сильные, но менее частые сигналы могут быть потеряны в простом объеме более слабых. Это не проблема, пока сигналы коррелируют положительно, но в случае отсутствия корреляции или отрицательной корреляции все меняется, скажем, люди нажимают на дорогие товары, но покупают доступные. Одним из способов противодействия подобному является нормализация по амплитуде каждого сигнала [17]; другой способ – рассмотрение отдельных целевых функций для каждого сигнала, а затем обеспечение их адекватного представления во время интерполяции.

Измерение успеха системы и информирование системы об этих измерениях создает петлю обратной связи. Это парадигма обучения с подкреплением, в которой система выполняет действие и получает обратную связь для обновления своего внутреннего состояния в сторону максимизации (или минимизации) целевой функции. Некоторые критерии можно измерить сразу после отображения результатов поиска, для других же требуется больше времени. Клики, переформулировки и отказы определяют быструю петлю обратной связи, которая длится от нескольких секунд до часов. В электронной коммерции существуют и более длинные циклы обратной связи, при которых обратная связь возникает уже после того, как система показала результаты пользователю. Приведем некоторые примеры: добавление товаров в корзину и проверка ее позже, добавление товаров в список пожеланий и проверка их на определенных событиях (например, Рождество, дни рождения) или покупка нескольких размеров товара, чтобы вернуть все, кроме подходящего. Удаление товаров из корзины или списка желаний, а также возврат товара – интересные типы отсроченной обратной связи, потому что задержка может составлять недели. Такие отсрочки затрудняют отслеживание и добавление отзывов к определенному рейтингу, поскольку для этого требуется сохранение записей рейтингов в течение длительных периодов времени. Так, остается открытым вопрос, как спроектировать системы, способные технически осуществимым образом фиксировать петли обратной связи с задержкой. Отметим, хорошие априорные значения для значений функций и бизнес-логики необходимы для преодоления проблем с отсутствием поведенческих данных для новых клиентов, называемых «холодным стартом».

Обратная связь перечисленных выше типов может быть использована для Learning to Match для оптимизации отдельных ранжировщиков или Learning to Rank (LtR) для оптимизации комбинаций отдельных ранжировщиков. Сообщество все чаще рассматривает возможность использования для этой цели исторических данных обратной связи, чтобы избежать вмешательств, негативно влияющих на качество обслуживания клиентов. Вопрос, как эффективно использовать широкий спектр сигналов обратной связи с сильно различающимися временными масштабами для фактического LTR, остается предметом исследований.

Поиск товара требует понимания характеристик документов и задачи поиска. Искусственные нейронные сети являются привлекательным решением, поскольку они могут получить такое понимание из необработанного текста документа и обучающих данных. Большинство нейронных методов фокусируются на изучении закономерностей релевантности запроса к документу, то есть знаний о задаче поиска. Однако изучение только шаблонов релевантности требует больших объемов обучающих данных и, тем не менее, все еще недостаточно хорошо обобщается для конечных запросов или новых доменов поиска. Эти проблемы формируют необходимость в предварительно обученных моделях понимания текста общего назначения. Предварительно обученные словесные представления, такие как fastText [18], широко используются в нейронных сетях для Information Retrieval (IR). Они изучаются на основе сочетания слов в большом корпусе, дающем информацию о синонимах и родственных словах. Но совпадение слов – это только поверхностное понимание текста с помощью набора слов. Недавно все мы стали свидетелями быстрого прогресса в понимании текста благодаря внедрению предварительно обученных нейронных языковых моделей, таких как BERT [19]. В отличие от традиционных эмбеддингов слов, они контекстуальны, представление слова является функцией всего входного текста, при этом учитываются зависимости слов и структуры предложений. Модели предварительно обучаются на большом количестве документов, поэтому контекстуальные представления кодируют общие языковые шаблоны. Контекстуальные нейронные языковые модели превзошли традиционные методы эмбеддингов слов в различных NLP задачах. Более глубокое понимание текста контекстуальными нейронными языковыми моделями открывает новые

возможности для IR, в том числе по использованию BERT (двунаправленных кодирующих представлений от Transformers) для поиска документов ad-hoc.

BERT – это современная нейросетевая модель, в значительной степени подходящая для задач поиска. BERT обучен предсказывать взаимосвязь между двумя фрагментами текста (обычно предложениями). Его архитектура, основанная на внимании, моделирует локальные взаимодействия слов в тексте1 со словами в тексте2. Что можно рассматривать как нейронную модель ранжирования, построенную на взаимодействии. Таким образом, требуется минимальное архитектурное проектирование, специфичное для поиска. В многочисленных исследованиях изучается влияние понимания языка BERT на поиск документов ad-hoc. Рассматриваются модели BERT на двух наборах данных для поиска ad-hoc с различными характеристиками. Эксперименты показывают, что тонкая настройка предварительно обученных моделей BERT с ограниченным объемом поисковых данных может обеспечить лучшую производительность, чем строгие исходные данные. В отличие от наблюдений, полученных с помощью традиционных моделей поиска, с помощью BERT более длинные запросы на естественном языке способны значительно превосходить запросы по коротким ключевым словам. Дальнейший анализ показывает, что стоп-слова и знаки препинания, которые часто игнорируются традиционными подходами к IR, играют ключевую роль в понимании запросов на естественном языке путем определения грамматических структур и зависимостей слов. Наконец, расширение BERT знаниями о поиске по большому журналу поиска позволяет получить предварительно обученную модель, оснащенную знаниями как о понимании текста, так и о поисковой задаче, что выгодно для связанной задачи поиска, где помеченные данные ограничены..

III. Эксперимент

ля эмпирического исследования взяты данные журналов поисковых запросов и текстовые представления карточек товаров электронной торговой Интернет-площадки. Далее создано несколько непубличных наборов данных для проверки гипотез.

Гипотеза: текстовые представления запроса и товара, купленного после введения запроса, зависимы. Эта зависимость может быть смоделирована.

Интуиция: в запросе содержится предмет, который клиент намеревается приобрести, в описании товара также содержится предмет. Существует связь «предмет в запросе» – «предмет в название товара», и эту связь возможно выявить на основании текстовых признаков.

В целях проверки гипотез использованы подходы на основе одно- и двух-кодирующих архитектур. Для одиночного кодировщика (single encoder) выбраны следующие методы:

- BM25 – разряженное кодирование текстовых признаков из карточек товаров,
- fastText – кодирование путем создания модели текста дистрибутивной семантики, обученной на публичном корпусе Википедии [18],
- BERT – кодирование с помощью модели текста на основе трансформера с обучением на непубличном наборе данных из текстового представления карточек товаров [19].

Также рассмотрены архитектуры на основе двух кодировщиков:

- DSSM (MLP) – двойной кодировщик, в качестве нейросетевых архитектур для кодировщиков использованы Multi-layer Perceptron [20].
- DSSM (CNN) – двойной кодировщик, в качестве нейросетевых архитектур для кодировщиков использованы Multi-layer Perceptron для текста запросов и CNN (ResNet) для картинок с изображением товара.

Для набора данных выбран 1 млн записей «Поисковый запрос» – «Товар (купленный)». Важно учитывать, что запросы бывают разной длины (в символах). Релевантность поисковой выдачи выше для коротких запросов и, соответственно, ниже для длинных. В качестве текстового представления карточки товара выбраны: название товара, описание товара, характеристики товара (бренд, цвет, вес и другое).

Для качественного понимания разницы в работе различных архитектур проведен эксперимент, в котором было закодировано 100 тысяч запросов и 100 млн карточек товаров. На рисунке 1 приведены распределения релевантности на основе косинусной близости для различных архитектур кодировщиков.

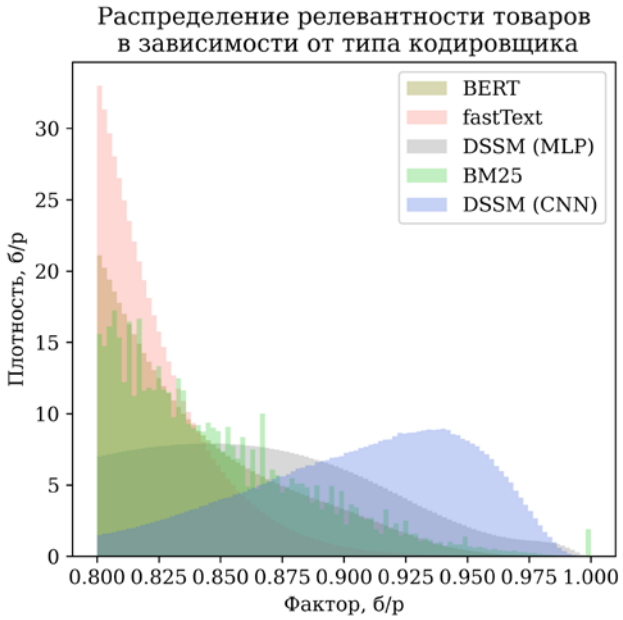

Рисунок 1 — Распределения релевантности запросов и товаров, полученные из разных типов кодировщиков

Из зависимостей на рисунке 1 видно, что модель дистрибутивной семантики текста (fastText) и модель текста на основе трансформеров (BERT) различаются не значительно. Большие значения релевантности в случае обучения на текстовых представлениях карточек

товаров для BERT вызваны смещением (bias) в текстах карточек товаров по сравнению с текстом в википедии, на котором обучен кодировщик основанный на fastText. Важно отметить, кодировщик на основе DSSM имеет еще более сильное смещение в силу того, что позволяет моделировать поведенческие связи между запросом и карточной товара. Для обучения кодировщика на основе DSSM использован подход обучения с учителем, в отличии от одиночных кодировщиков, обученных без учителя. Также следует подчеркнуть, что в случае с использованием различных модальностей представления товара, например, картинок, смещение наибольшее из рассмотренных типов кодировщиков.

Дальнейшее экспериментальное сравнение влияния архитектур кодировщиков на релевантность проведено на основании метрик эффективности поиска – Precision, Recall и NDCG.

Для вычисления метрик эффективности поиска использована парадигма Карнфилда: предполагается, что мы ищем набор товаров, которые проиндексированы различными системами поиска и сравниваем, насколько хорошо эти системы работают.

В таком случае метрики полноты (Precision, P) и точности (Recall, R) будут определяться по формулам, приведенным на рисунке 2.

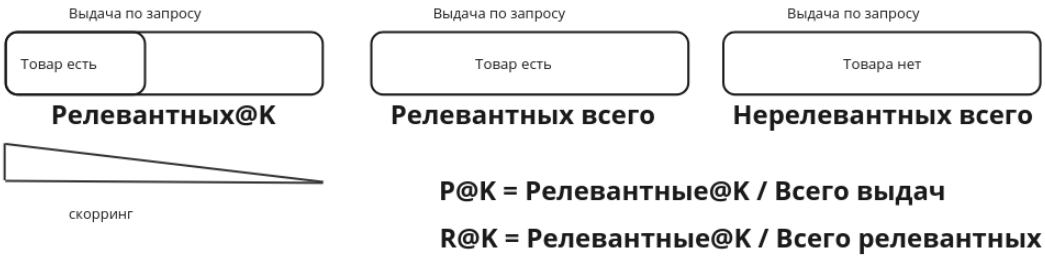

Рисунок 2– Формулы метрик эффективности поисковых выдач по товарам

На электронной торговой Интернет-площадке ранг не очень равномерно уменьшается, так как товары представлены в виде таблицы со строкой 5-7 элементов. «Первым рангом» обладает первый ряд с товарами из выдачи. Вероятность просмотра падает ступенчато, а не гладко.

Далее согласно схеме, изображенной на рисунке 2, получены метрики эффективности поиска: Точность@K (таблица 1) и Полнота@K (таблица 2).

Таблица 1 – Точность (precision) поисковой выдачи для первых @K товаров

	P@1	P@7	P@12	P@49	P@100
BERT	36.38%	59.14%	65.08%	78.36%	83.86%
DSSM(MLP)	18.15%	40.95%	48.41%	68.82%	77.91%
DSSM(CNN)	0.45%	2.26%	3.36%	9.12%	14.70%
BM25	42.88%	65.22%	70.28%	81.12%	85.23%
fastText (Wiki)	36.34%	54.18%	58.81%	69.41%	74.13%

Таблица 2– Полнота (recall) поисковой выдачи для первых @K товаров

	R@1	R@7	R@12	R@49
BERT	43.39%	70.53%	77.61%	93.44%
DSSM(MLP)	23.30%	52.56%	62.13%	88.33%
DSSM(CNN)	3.06%	15.36%	22.85%	62.02%
BM25	50.31%	76.52%	82.46%	95.18%
fastText (Wiki)	49.02%	73.09%	79.33%	93.63%

Следует отметить, что лексический поиск на основе текстовых признаков (BM25) продемонстрировал самые высокие значения метрик в силу самого метода вычисления метрик, использованного автором, и скорее может считаться baseline для данного исследования. Сам факт признания выдачи релевантной определяется по текстовому совпадению названия товара, купленного по данному запросу, с названием товара из поисковой выдачи, полученной различными методами. Интерес

настоящего исследования представляют относительные позиции методов извлечения признаков (features) с использованием искусственных нейронных сетей глубокого обучения. Так, наиболее эффективным показал себя fastText, предобученный на корпусе документов Wikipedia. Этому эмпирическому наблюдению можно дать следующие объяснения. Во-первых, субсловарная токенизация, использованная в fastText, хорошо справляется с проблемой неизвестных при обучении слов (OOV). Во-вторых, Wikipedia представляет хороший корпус для обучения русскому языку. Обучение BERT на «золотых» (наиболее качественных по тексту) текстовых представлениях товаров имеет смещения в сторону сленга продавцов. В-третьих, сам подход к дистрибутивной семантике, реализованный в fastText, наиболее отвечает требованиям к синонимичности в поисковых запросах. В-четвертых, fastText устойчив к опечаткам, доля которых в поисковых запросах составляет более 15 %. С другой стороны, DSSM в обоих вариантах не показал высоких значений метрик по совершенно понятным причинам. При обучении DSSM моделей учитывается именно факт покупки, а не релевантности. Таким образом, DSSM в явном виде имеет смещение на покупаемые товары, а не на товары, соответствующие поисковому запросу (релевантные).

Для вычисления метрики NDCG@K автором сделано следующее допущение: товар, который соответствует запросу, должен находиться на первой позиции в поисковой выдаче.

Таблица 3 — Релевантность (NDCG) поисковой выдачи для первых @K товаров

	NDCG@1	NDCG@7	NDCG@12	NDCG@49	NDCG@100
BERT	36.38%	45.25%	46.64%	49.03%	49.81%
DSSM(MLP)	18.15%	26.80%	28.54%	32.21%	33.49%
DSSM(CNN)	0.45%	1.11%	1.37%	2.38%	3.15%
BM25	42.88%	51.67%	52.86%	54.82%	55.40%
fastText (Wiki)	36.34%	43.35%	44.44%	46.34%	47.01%

Полученные результаты по NDCG демонстрируют, что BERT является несколько лучшим кодировщиком, чем другие кодировщики на основе сетей глубокого обучения. Отметим, BERT был обучен с использованием CUDA и требовал значительно большего времени обучения, по сравнению с другими кодировщиками.

IV. ЗАКЛЮЧЕНИЕ

Нейронные сети предоставляют новые возможности для автоматического изучения сложных языковых паттернов и связей между запросами и документами. Нейронные модели достигли многообещающих результатов в изучении закономерностей релевантности запроса к документу, но было проведено недостаточно исследований по пониманию текстового содержимого запроса или документа. В данной статье исследовано использование недавно предложенной контекстуальной нейронной языковой модели BERT для обеспечения более глубокого понимания текста в IR. Экспериментальные результаты демонстрируют, что контекстуальные текстовые представления от BERT более эффективны, чем традиционные компактные векторные представления слов. По сравнению с моделями поиска по набору слов, контекстная языковая модель может лучше использовать языковые структуры, внося значительные улучшения в запросы, написанные на естественных языках. Сочетание способности понимать текст со знаниями в области поиска приводит к созданию усовершенствованной, предварительно обученной модели BERT, которая может принести пользу смежным задачам поиска, в которых обучающие данные ограничены.

Семантические сигналы, извлеченные из пары «запрос — купленный товар», представляют, наряду с текстовыми (лексическими) и текстовыми семантическими признаками, комплементарную основу для обучения ранжировщиков. В случае, если семантический фактор добавляет полноту, можно считать, что текстовый фактор для этого товара не существенен. А в случае, если оба фактора (лексический и семантический) могут пересортировывать выдачу, то в данной статье приведены количественные оценки влияния отдельных признаков на NDCG. Наконец, сигналы ранжирования открывают возможность разработки многоцелевых функций, циклов обратной связи с помощью обучения ранжированию (Learning to Rank, LtR).

Библиография

[1] Yang Y. Query-aware tip generation for vertical search // Proceedings of the 29th ACM International Conference on Information & Knowledge Management. – 2020. – P. 2893-2900.

[2] Ibrahim, Osman Ali Sadek and Eman M. G. Younis. Hybrid online–offline learning to rank using simulated annealing strategy based on dependent click model // Knowledge and Information Systems. – 2022. – Vol. 64. – P. 2833-2847.

[3] Lyu, Zequn Deep Match to Rank Model for Personalized Click-Through Rate Prediction // AAAI Conference on Artificial Intelligence. – 2020.

[4] Zhai, Jiaqi Revisiting Neural Retrieval on Accelerators // ArXiv abs/2306.04039. – 2023.

[5] C. Rudin. The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list // The Journal of Machine Learning Research. – 2009. – Vol. 10. – P. 2233–2271.

[6] Krauth, K., Dean, S., Zhao, A., Guo, W., Curmei, M., Recht, B., & Jordan, M. I. Do offline metrics predict online performance in recommender systems? // arXiv preprint arXiv:2011.07931. – 2020.

[7] Curmei, M., Haupt, A. A., Recht, B., & Hadfield-Menell, D. Towards psychologically-grounded dynamic preference models // In Proceedings of the 16th ACM Conference on Recommender Systems. – 2022. – P. 35-48.

[8] Patro, G. K., Porcaro, L., Mitchell, L., Zhang, Q., Zehlike, M., & Garg, N. Fair ranking: a critical review, challenges, and future directions // In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. – 2022 (June). – P. 1929-1942.

[9] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan Expected reciprocal rank for graded relevance // In Proceedings of the 18th ACM conference on Information and knowledge management. ACM. – 2009. – P. 621–630.

[10] Yun Zhou and W Bruce Cro. Query performance prediction in web search environments // In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM. – 2007. – P. 543–550.

[11] Cronen-Townsend, Steve [et al] Predicting query performance // Annual International ACM SIGIR Conference on Research and Development in Information Retrieva. – 2002.

[12] Järvelin, Kalervo and Jaana Kekäläinen Cumulated gain-based evaluation of IR techniques // ACM Trans. Inf. Syst. – 2002. – Vol. 20. – P. 422-446.

[13] Evangelos Kanoulas and Javed A. Aslam. Empirical justification of the gain and discount. – 2009.

[14] Carmel, D., Haramaty, E., Lazerson, A., & Lewin-Eytan, L. Multi-objective ranking optimization for product search using stochastic label aggregation // In Proceedings of The Web Conference. – 2020 (April). – P. 373-383.

[15] Hong, Liangjie, and Mounia Lalmas Tutorial on online user engagement: Metrics and optimization // Companion Proceedings of The 2019 World Wide Web Conference. – 2019.

[16] Reddy, Chandan K. Shopping queries dataset: A large-scale ESCI benchmark for improving product search // arXiv preprint arXiv:2206.06588. – 2022.

[17] Tsagkias, Manos, Wouter Weerkamp, and Maarten De Rijke News comments: Exploring, modeling, and online prediction // Advances in Information Retrieval: 32nd European Conference on IR Research, ECIR. – Milton Keynes, UK, March 28-31, 2010.

[18] Bojanowski P. Enriching word vectors with subword information // Transactions of the association for computational linguistics. – 2017. – Vol. 5. – P. 135-146.

[19] Rogers A. RuSentiment: An enriched sentiment analysis dataset for social media in Russian // Proceedings of the 27th international conference on computational linguistics. – 2018. – P. 755-763.

[20] Huang P. S. 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.

Статья получена 19 сентября 2023. Ф.В.Краснов,
Исследовательский центр ООО "ВБ СК" на базе Инновационного Центра Сколково.
krasnov.fedor2@wb.ru, http://orcid.org/0000-0002-9881-7371

Investigation of the influence of textual representation of goods by means of models using artificial neural networks of deep learning on the relevance of the search for goods on the electronic trading Internet platform

F.V. Krasnov

Abstract — With the growing popularity of online shopping, academic research in the field of e-commerce is becoming increasingly relevant. Nevertheless, significant research problems remain, starting with classical search problems in e-commerce, for example, the comparison of text queries with multimodal representations of goods, optimization of ranking for bilateral Internet trading platforms, taking into account the interaction of buyer and seller with recommendation and search engines. These areas of research are important for understanding customer behavior, stimulating engagement, and increasing conversions.

Product data plays a key role in search and recommendations. Services in which there are several suppliers (marketplaces) on the electronic trading Internet platform demonstrate high dynamics in terms of the quality and consistency of content, fraud detection and pricing. Inventory updates are often accompanied by service level agreements that guarantee that changes will be made in accordance with customer requirements within a strictly defined time frame. Finally, the goods themselves can be considered as multimodal documents, and the success of the search depends on the compliance of the client's intentions with all aspects of the goods. For the sake of specificity, we will use the term "documents" for the elements returned by the search query, although the returned elements may be more general (for example, multimedia elements).

When searching on an online electronic trading platform, the indexed products that buyers are looking for are combinations of images, videos and unstructured text (names, descriptions and reviews) and structured data (price, brand, ratings, seller's location, and logistics). This combination of data opens up prospects for research, including improving the extraction of data from a variety of sources using signals from different types of data, such as providing more detailed color information and using image similarity for recommendations. and also as a way for customers to create queries in a search engine. Data is the key to providing high-quality search and recommendations that meet the needs of customers and businesses.

In this article, the author explored the possibilities of several approaches to extracting useful signals from various sources of product information to improve the customer experience in the field of e-commerce.

Key words — *NDCG, DNN, transformers, BERT, e-commerce, Information Retrieval, IR.*

REFERENCES

[1] Yang Y. Query-aware tip generation for vertical search // Proceedings of the 29th ACM International Conference on Information & Knowledge Management. – 2020. – P. 2893-2900.

[2] Ibrahim, Osman Ali Sadek and Eman M. G. Younis. Hybrid online–offline learning to rank using simulated annealing strategy based on dependent click model // Knowledge and Information Systems. – 2022. – Vol. 64. – P. 2833-2847.

[3] Lyu, Zequn Deep Match to Rank Model for Personalized Click-Through Rate Prediction // AAAI Conference on Artificial Intelligence. – 2020.

[4] Zhai, Jiaqi Revisiting Neural Retrieval on Accelerators // ArXiv abs/2306.04039. – 2023.

[5] C. Rudin. The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list // The Journal of Machine Learning Research. – 2009. – Vol. 10. – P. 2233–2271.

[6] Krauth, K., Dean, S., Zhao, A., Guo, W., Curmei, M., Recht, B., & Jordan, M. I. Do offline metrics predict online performance in recommender systems? // arXiv preprint arXiv:2011.07931. – 2020.

[7] Curmei, M., Haupt, A. A., Recht, B., & Hadfield-Menell, D. Towards psychologically-grounded dynamic preference models // In Proceedings of the 16th ACM Conference on Recommender Systems. – 2022. – P. 35-48.

[8] Patro, G. K., Porcaro, L., Mitchell, L., Zhang, Q., Zehlike, M., & Garg, N. Fair ranking: a critical review, challenges, and future directions // In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. – 2022 (June). – P. 1929-1942.

[9] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan Expected reciprocal rank for graded relevance // In Proceedings of the 18th ACM conference on Information and knowledge management. ACM. – 2009. – P. 621–630.
[10] Yun Zhou and W Bruce Cro. Query performance prediction in web search environments // In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM. – 2007. – P. 543–550.
[11] Cronen-Townsend, Steve [et al] Predicting query performance // Annual International ACM SIGIR Conference on Research and Development in Information Retrieva. – 2002.
[12] Järvelin, Kalervo and Jaana Kekäläinen Cumulated gain-based evaluation of IR techniques // ACM Trans. Inf. Syst. – 2002. – Vol. 20. – P. 422-446.
[13] Evangelos Kanoulas and Javed A. Aslam. Empirical justification of the gain and discount. – 2009.
[14] Carmel, D., Haramaty, E., Lazerson, A., & Lewin-Eytan, L. Multi-objective ranking optimization for product search using stochastic label aggregation // In Proceedings of The Web Conference. – 2020 (April). – P. 373-383.
[15] Hong, Liangjie, and Mounia Lalmas Tutorial on online user engagement: Metrics and optimization // Companion Proceedings of The 2019 World Wide Web Conference. – 2019.
[16] Reddy, Chandan K. Shopping queries dataset: A large-scale ESCI benchmark for improving product search // arXiv preprint arXiv:2206.06588. – 2022.
[17] Tsagkias, Manos, Wouter Weerkamp, and Maarten De Rijke News comments: Exploring, modeling, and online prediction // Advances in Information Retrieval: 32nd European Conference on IR Research, ECIR. – Milton Keynes, UK, March 28-31, 2010.
[18] Bojanowski P. Enriching word vectors with subword information // Transactions of the association for computational linguistics. – 2017. – Vol. 5. – P. 135-146.
[19] Rogers A. RuSentiment: An enriched sentiment analysis dataset for social media in Russian // Proceedings of the 27th international conference on computational linguistics. – 2018. – P. 755-763.
[20] Huang P. S. 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.