Федеральное государственное автономное образовательное
учреждение высшего профессионального образования
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

На правах рукописи

Сорокина Анна Николаевна

ОПТИМИЗАЦИЯ ПОКАЗОВ РЕКЛАМНЫХ ОБЪЯВЛЕНИЙ В ПОИСКОВЫХ ИНТЕРНЕТ-СИСТЕМАХ: РАЗРАБОТКА МЕТОДОЛОГИИ ПОДБОРА ПОРОГОВ ВХОДА В РЕКЛАМНЫЙ ПОКАЗ

Специальность 05.13.18 – «Математическое моделирование, численные методы и комплексы программ»

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель
д. ф.-м. н.
Цитович И.И.

Москва – 2015

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ.

В настоящий момент огромное количество людей пользуются Интернетом для поиска информации. Обычно Интернет-пользователь задает некоторый **запрос**, на который он хочет получить ответ, и задача поисковых систем – отвечать на эти запросы наилучшим образом. Чтобы иметь прибыль от поиска информации для пользователя, поисковик использует рекламные объявления, которые показываются на странице поисковой выдачи. Вообще говоря, не все рекламные объявления, отобранные по запросу, могут быть показаны из-за ограничений на странице поиска. Если пользователь кликает на то или иное объявление, происходит переход на рекламируемый сайт; этот переход называется **кликом**. В случае клика осуществляется **списание денег** со счета рекламодателя, причем общая сумма от списания денег со счетов рекламодателей оказывается существенной составляющей дохода поисковой компании. Рекламодатель для каждого объявления выставляет **ставку** (количество денег, которые рекламодатель готов заплатить за клик); тем самым происходит торг за показ его рекламного объявления на поиске.

Рекламные объявления могут показываться в разных местах поисковой страницы, в диссертационном исследовании рассматривается рекламный блок, который показывается над результатами поиска по запросу пользователя. Из-за своего расположения, этот рекламный блок является наиболее привилегированным местом для размещения рекламы, поэтому он приоритетен для рекламодателей. Так же рекламный блок над результатами поиска является самым прибыльным для поисковой системы: зачастую, для того чтобы попасть в него, рекламодателю необходимо поставить достаточно большую ставку (из-за конкуренции).

Целесообразность показа рекламного объявления над результатами поиска для поисковой системы определяется двумя критериями. Первый – это вероятность того, что пользователь кликнет на предъявленное объявление (кликабельность – $CTR, Click-Through-Rate$). Можно сказать, что это – степень эффективности показа объявления, его характеристика привлекательности для пользователя. Второй – это ставка (Bid), назначенная рекламодателем за показ его объявления в спец-размещении.

Ожидаемое количество денежных средств, списываемых со счета рекламодателя за клик пользователя по конкретному объявлению, может быть оценено как произведение вероятности клика на ставку, назначенную за объявление. Именно эта величина CPM ($Cost-Per-Million$) $= CTR \cdot Bid$ в настоящее время служит критерием, по которому отбираются кандидаты для показа над результатами поиска.

Таким образом, важно построить критерий показа рекламного объявления над результатами поиска с учётом CTR и CPM, так как это два ключевых показателя эффективности показов рекламных объявлений.

Актуальность проблемы

Оптимальный отбор объявлений для показа над результатами поиска – важная практическая задача для современных поисковых интернет-систем. Ее решение – в интересах и других сторон: рекламодателей и самих пользователей, поскольку первые нуждаются в привлечении потребителей, а вторые – в релевантной их запросу информации.

В научной литературе проблематика выбора наилучшего критерия показа рекламного объявления стала активно изучаться последние 10-15 лет. Над обозначенной проблемой работали многие учёные: Choul

W. L., Agarwal D. K., Broder A. Z., Ciaramita M., Lacerda A., Ghose A., Zhang W. V., Schroedl S., Lahaie S., Radlinski F., Chakrabarti D., Dembczynski K., Regelson M., Richardson M., Feng J., Joachims T., Granka L., Herbrich R., Liu T. Y., Zhu Y. Graepel T., Sheth A. и др. Как видно, проблема является актуальной и широко исследуется научным сообществом, однако ни в одной из представленных в литературе работ не уделяется достаточное внимание исследованию модели показов рекламных объявлений, выявлению критериев эффективности показов и соответствующих ограничений, математической постановке задачи оптимизации и её решению. Обусловлено это тем, что решение показа рекламного объявления на запрос пользователя требует учета большого количества факторов и переменных, которые между собой взаимосвязаны. Решение этой важной и сложной научно-практической проблемы требует применения современного математического аппарата, а именно – разработке математической модели показов рекламных объявлений, а также постановки и решения задачи оптимизации, что и было сделано в представленной работе.

Степень разработанности проблемы

В научной литературе проблематика выбора критерия показа рекламного объявления стала активно изучаться последние 10-15 лет. В литературе можно выделить три подхода к постановке задачи оптимизации рекламных показов.

Первый подход использует в качестве способа обучения алгоритма ранжирования рекламных объявлений оптимизацию **релевантности** рекламы [46], [52], [16]: их цель состоит в максимизации количества кликов. В случае если ранжирующая функция может точно предсказывать клик по рекламному объявлению и его релевантность, то этот подход также приводит к максимизации

дохода поисковой системы. Чтобы учесть суммарный доход поисковой системы, один подход объединяет в себе доход и релевантность в форме линейной комбинации [52], другой – умножение показателя релевантности на ставку рекламодателя [14]. Оба метода являются эвристическими, поэтому нахождение оптимальных параметров достаточно затруднительно. Когда настраиваемых параметров достаточно много, эвристический выбор этих параметров становится невозможным, т. к. вычислительная сложность увеличивается экспоненциально с ростом числа параметров.

Второй подход состоит в разделении ожидаемого дохода на две составляющие, каждую из которых стараются максимизировать: CTR и ставка рекламодателя. Предложено несколько алгоритмов для предсказания CTR объявления с высокой точностью [27], [55], [53], [32]. Предположением этого подхода, как указано в [32], является то, что существует собственная релевантность объявления, которая не зависит от контекста запроса пользователя. Объявление с высоким значением CTR может быть и неревантным конкретному запросу пользователя.

Механизм, изложенный в работе [46], дает результаты только в случае высокой точности предсказания релевантности объявления. Вторая [52] и третья [16] работы используют двух-шаговую оптимизацию дохода вместо релевантности. Несмотря на то, что в этих работах действительно есть определённое увеличение доходности, релевантность значительно падает.

Третий подход состоит в следующем: как обсуждалось в [43], [35], существуют два важных фактора, от которых зависит решение пользователя о том кликнуть на определённое объявление или нет. Во-первых, это – **позиционность** (зависимость вероятности клика от позиции, на котором показывается объявление), которое приводит к

увеличению числа кликов на рекламные объявления, размещённые на верхних позициях. Другая состоит в том, что на вероятность клика пользователя влияет не только абсолютная релевантность конкретного объявления, но и общее **качество других объявлений** в рекламном хите. Многие авторы строят критерий показа исходя именно из этих двух факторов и достигают достаточно хороших результатов [38], [43], [49]. В работе Чжу [67] предлагается два подхода к сочетанию прибыли и релевантности рекламных объявлений, связанных с построением сложных целевых функций.

Когда оптимизационная задача поставлена, необходимо зафиксировать **критерий показа** объявлений. **Классическим критерием показа** объявлений над результатами поиска является $CPM = Bid \cdot CTR$.

Одна из крупнейших поисковых компаний Yahoo! сортирует рекламные объявления в **зависимости от точности соответствия фраз запроса и купленной фразы**, и только после этого по ставке, соответствующей рекламному объявлению [36]. Система показов интернет-рекламы Google AdWords в 2002 году изменила критерий показа рекламных объявлений со ставки за клик на **ставку за клик умноженную на кликабельность.** Yahoo! начал вычислять метрику *Click Index* приводя **позиционный наблюдаемый** *CTR* **к стандартному**. Этот стандартный *CTR* показывается в интерфейсе рекламодателя как их «истинный» *CTR*, но ранжирование происходит уже по другому критерию.

Таким образом, при сортировке по CPM отбор объявления зависит только от вероятной выгоды его показа для поисковой системы, что является не совсем корректным по отношению к пользователям и рекламодателям.

В научной литературе встречается достаточно большое количество работ с различными критериями показа объявлений в рекламном блоке над результатами поиска.

Первым из критериев показа, описанным в [7], является $Bid \cdot quality + v$, где $quality$ - **качество** объявления (в нашем случае это предсказываемый CTR), а $\boldsymbol{v}$ - количественная **мера полезности** объявления для поисковой системы в качестве обучающего материала и дополнительной информации о качестве (вообще говоря это некоторое начальное предсказание CTR для новых баннеров). При выборе v нужно следить за балансом между показами новых и старых объявлений.

В продолжение этого критерия существует такая схема отбора объявлений, как $\boldsymbol{bid \cdot quality + v}$, где $\boldsymbol{v = c \cdot bid \cdot var(quality)}$, а c – константа, показывающая на сколько полезно обучение на данном объявлении, т.е. **баланс между долгосрочным и краткосрочным показом объявления**. Данный метод интересен тем, что начальный прогноз качества объявления не является одной константой, а зависит от разброса значений этого прогноза и его ставки, т.е. отбирается то объявление на показ, которое принесёт поисковой системе больший доход.

Вторым направлением для выбора критерия показа объявлений является **релевантность** – степень соответствия содержания рекламного объявления запросу пользователя (семантическое соответствие поискового запроса и текста объявления) [14], [21], [46]. Но отбор посредством релевантности учитывает только интересы пользователя: рекламодатель практически не может напрямую повлиять на отбор его объявления на показ, т.к. его ставка не входит в критерий показа рекламных объявлений над результатами поиска. Так же и поисковая интернет-система имеет достаточно ограниченное

влияние на свои основные показатели, такие как суммарные клики, среднюю кликабельность и доход: отбор объявлений зависит только от запроса пользователя и от того, насколько корректно составлен текст объявления и оптимизирован сайт рекламодателя под текущий запрос.

Третьим подходом является выбор в качестве характеристик, от которых зависит показ объявлений, следующих: на сколько фраза, по которой показывается объявление, является **широко-тематической** или **узко-тематической**. Тут возникает проблема баланса между фразами: зачастую бывает достаточно сложно определить тематическую принадлежность фраз; для этого нужно собирать и хранить большое количество информации, а так же обновлять её по мере накопления статистики.

Следующей характеристикой для является **информация о продавце**, а так же информация о **бренде**, представляемом объявлением [33]. Однако, тут есть сложность в достоверности данных для разных пользователей, а также изменение предпочтений и вкусов. Также не всегда представляется возможным собрать достаточную статистику по всем рекламодателям и брендам, а для тех, кто впервые встретился пользователю, информации нет никакой и как их показывать остаётся под вопросом.

Четвёртый подход среди работ, посвящённых критерию показа объявлений в рекламном блоке, основан на большом внимании к **позиционным моделям** показа рекламы [66]. Было выявлено, что позиция, на которой было показано объявление, влияет на его кликабельность, следовательно, важно её учитывать при отборе объявлений на показ. Также для разных объявлений показ на той или иной позиции может иметь различный эффект, важно учесть совместный показ группы объявлений, причём каждого на своей собственной позиции. Позиционные эффекты будут рассмотрены в

данной работе как продолжение одного из методов подбора критерия показа над результатами поиска.

Пятый подход состоит в том, что совместно с позиционной моделью часто рассматривается модель, зависящая от **пользователя**, т.е. в критерий показа добавляется элемент, зависящий от характеристик пользователя. Например, в [57] критерий показа над результатами поиска имеет вид: $CTR_i \cdot pos_i \cdot ucp \cdot bid_i \geq \theta_i$, где добавляется мультипликативный множитель ucp – персональный признак, обозначающий на сколько данный пользователь расположен кликать на рекламу в данный момент по сравнению со среднестатистическим пользователем.

Шестым в качестве критерия показа объявления в рекламный блок рассматривались критерии вида $bid^k \cdot CTR^l$. Данный вид критерия даёт достаточно большую степень свободы при отборе объявлений. Однако, ограниченность данного метода заключается в том, что ставка и предсказанная кликабельность берутся именно в произведении одно на другое. То есть $bid^k \cdot CTR^l$ при любых k и l – некоторая вероятность списания ставки за клик по объявлению, мы можем только усиливать зависимость критерия показа либо от ставки данного объявления, либо от предсказанного CTR объявления.

Как видно из представленного обзора, тема критерия показа рекламных объявлений над результатами поиска актуальна и широко исследуется научным сообществом. Также важно заметить, что некоторые авторы ставят оптимизационные задачи, однако в рассмотренных работах есть некоторые ограничения и алгоритмы больше эвристические и подходят только лишь к конкретным постановкам задачи оптимизации.

Ни в одной из представленных в литературе работ не встречается разработка модели показов рекламных объявлений, её математическое

описание и соответствующая математическая постановка задачи оптимизации рекламных показов и её решение.

Цель и задачи исследования

Цель работы – математическая постановка задачи оптимизации и разработка алгоритма её решения.

Задачи диссертационного исследования:

1. Исследовать модель показов рекламных объявлений, провести анализ существующих подходов к отбору объявлений для показа, их ограничений и недостатков.
2. Разработать новую методику решения проблемы размещения рекламных объявлений над результатами поиска, а также на основе разработанной методики поставить соответствующую математическую задачу оптимизации.
3. Решить задачу оптимизации и получить алгоритм отбора объявлений в рекламный показ. С помощью критериев, полученных с помощью алгоритма, научиться работать с новыми поисковыми запросами.
4. Разработать программные средства для реализации полученного алгоритма отбора и подбора его параметров.
5. Провести экспериментальное тестирование полученного алгоритма на базе поисковой интернет-системы «Яндекс», сделать выводы об эффективности выработанной методики, и полученного на её основе алгоритма, а также о наиболее перспективных направлениях её дальнейшей доработки и усовершенствования.

Предмет и объект исследования

Объект исследования: система показов рекламных объявлений на странице поисковой выдачи.

Предмет исследования: критерий показа рекламных объявлений в ответ на поисковый запрос пользователя в сети Интернет

Научная новизна и практическая ценность.

Научная новизна диссертационного исследования заключается в следующем:

- Разработана новая математическая модель показов рекламных объявлений над результатами поиска.
- Новая математическая модель учитывает дополнительные ограничения системы показов рекламных объявлений, а также максимизирует выявленный показатель эффективности показов.
- По-новому сформулированной математической задаче оптимизации с помощью релаксации предложено решение с помощью метода множителей Лагранжа, характерной особенностью которого является декомпозиция целевой функции.
- Получен новый критерий показа объявлений в рекламном блоке над результатами поиска.

Положения, выдвигаемые на защиту.

1. Разработана математическая модель показов рекламы, учитывающая дополнительные ограничения системы показов рекламных объявлений, а также максимизирующая выявленный показатель эффективности показов.
2. Разработанная методология математической постановки задачи оптимизации показов рекламных объявлений позволяет эффективно учитывать всю совокупность ключевых факторов, влияющих на основные показатели и характеристики системы показов рекламных объявлений.

3. Предложенное решение задачи оптимизации позволило получить новый вид критерия показа рекламных объявлений в рекламном блоке над результатами поиска, характерной особенностью которого является независимость от других объявлений.
4. Построен алгоритм нахождения параметров показа для выбранной задачи оптимизации рекламных показов.
5. Применение нового вида критерия показало повышение эффективности и производительности системы рекламных показов по итогам апробации в поисковой системе «Яндекс».

Практическая ценность состоит в том, что на основе проведенных исследований разработан алгоритм подбора параметров критерия показа рекламных объявлений над результатами поиска в современной поисковой интернет-системе (на примере системы Яндекс.Директ). Реализация данного подхода позволила повысить эффективность работы системы при заданных ограничениях. Разработанные теоретические положения и методики могут быть использованы и для других постановок задач оптимизации и других видах ограничений, в том числе позволяет использовать при отборе дополнительную информацию про объявления, например такую как релевантность запросу или конверсионность (действие пользователя на сайте рекламодателя).

Внедрение результатов диссертационного исследования в учебный процесс в рамках курсов лекций и практических занятий по дисциплинам «Современные методы анализа данных» и «Научный семинар «Анализ интернет-данных» позволило повысить теоретический и практический уровень знаний студентов в области методов анализа интернет-данных на примере изучения механизмов показа рекламы в поисковых интернет-системах.

Апробация и внедрение результатов исследования.

Основные результаты работы представлялись и получили одобрение на:

- Международной конференции International World Wide Web Conferences 2013 (13-17 мая 2013, Рио-Де-Жанейро, Бразилия) с постером «Optimization of ads allocation in sponsored search».
- Конференции ACM SIGKDD 2012 (12-16 августа 2012, Пекин, Китай) с докладом на семинаре The Sixth International Workshop on Data Mining for Online Advertising and Internet Economy на тему «Using boosted trees for click-through rate prediction for sponsored search».
- Научных семинарах базовой кафедры Яндекс НИУ ВШЭ.

По теме диссертационной работы опубликовано пять научных работ [1], [2], [5], [19], [61], включая две статьи в изданиях из списка изданий, рекомендованных ВАК РФ [1], [2]. В опубликованных статьях отражено основное содержание диссертационной работы.

Разработанный в диссертации алгоритм был реализован в онлайн-системе показов рекламы компании «Яндекс». Для подбора оптимальных значений параметров критерия показа был ***реализован комплекс программ***, включающий в себя следующие компоненты:

1) **QueryPool.py** – выбор из лога показов рекламных объявлений случайного ***набора запросов*** необходимого размера. На вход программе даются логи рекламных показов: временная последовательность запросов пользователей, на которые было показано хотя бы одно рекламное объявление. Пользовательские запросы нужны для того, чтобы собрать ***обучающий материал*** для подбора оптимальных параметров.
2) **GetCandidatesForQuery.py** – для каждого из запросов из базы данных рекламных объявлений выбираются ***соответствующие***

ему объявления. Для каждого из запросов, которые были собраны программой **QueryPool.py** , производятся следующие действия:

- Из запроса выделяются ключевые слова – слова, несущие основную смысловую нагрузку.
- Из ключевых слов (если их несколько) составляются ключевые фразы.
- По ключевым фразам из всего набора рекламных объявлений выбираются именно те, которые торгуются по соответствующим фразам из запроса. Таким образом, для каждого запроса отбираются соответствующие ему кандидаты на показ в рекламном блоке над результатами поиска.
- Для каждого объявления-кандидата известна его ставка Bid и вычисляется предсказание вероятности клика CTR.

В результате работы программы получается ***обучающий материал для подбора оптимальных значений параметров порогов.***

3) **OptimalThresholdParameters.py** – реализация ***алгоритма подбора оптимальных значений*** параметров функции порога в зависимости от поставленной оптимизационной задачи. На вход программа получает:
   a. Набор запросов с баннерами-кандидатами, для каждого из которых известны значения Bid и CTR, которые были получены предыдущей программой **GetCandidatesForQuery.py**
   b. Критерий, который необходимо максимизировать
   c. Ограничения поисковой системы.

   После того, как задана конкретная задача оптимизации с ограничениями, выполняется поиск соответствующих значений параметров.

Реализация комплекса программ была выполнена на языке программирования Python. Из-за больших объёмов данных (логов

показов рекламы, обучающего набора запросов и баннеров-кандидатов на показ) возникла необходимость использования ***распределённых вычислений*** на MapReduce: вычисления оптимальных параметров происходят на порядок быстрее.

Структура диссертационного исследования.

Основной текст диссертации изложен на 138 страницах, состоит из введения, четырёх глав, заключения, списка использованной литературы, состоящего из 74 наименований, и приложения, включающего в себя акты о внедрении.

В **первой главе** строится математическая модель показов рекламы и на основе анализа научной литературы обосновывается необходимость выработки новой модели, которая будет учитывать ограничения системы показов рекламы, а так же максимизировать показатель эффективности рекламных показов.

Также рассматривается структура рекламного блока и порядок его функционирования. В этой главе также говорится о развитии интернет-рекламы, выделении проблемы её показов перед результатами поиска. В конце главы приводятся обзор научных трудов на данную тематику, и вытекающая постановка задачи исследования.

Вторая глава посвящена математическому описанию модели показов рекламы, а также методологическому обоснованию постановки задачи оптимизации и построению нового алгоритма подбора параметров критерия показа рекламных объявлений над результатам поиска. Вначале ставится математическая постановка задачи оптимизации рекламных показов, затем приводится метод формирования критерия показа рекламных объявлений на основе решения задачи максимизации средней кликабельности рекламных объявлений при ограничении на общий доход системы и на покрытие

поисковых запросов рекламными блоками над результатами поиска с помощью метода множителей Лагранжа. После того как оптимизационная задача была решена, представлен формальный алгоритм подбора параметров критерия показа, работающий на реальных данных, полученных из работающей интернет-системы. Представлена работа полученного критерия показа для новых запросов. Далее во второй главе представлена модификация исходного (базового) алгоритма, которая значительно упрощает его работу и интерпретацию результатов его работы. В конце второй главы представлены результаты тестирования улучшенного алгоритма для каждого из его этапов.

В **третьей главе** диссертационного исследования рассматривается обобщение алгоритма подбора параметров критерия показа на предсказанную вероятность клика, зависящую от позиции, на которую попадёт рекламное объявление. В начале главы представлены основные положения по учёту позиционного эффекта. С учётом этих положений представлена математическая постановка задачи оптимизации рекламных показов и её решение с помощью метода множителей Лагранжа. В ходе оптимизации решено использовать задачу о назначениях и венгерский алгоритм. После того, как новый критерий показа получен, проводится численный эксперимент на симуляционных модельных данных: сравнивается новый алгоритм, учитывающий позиционный эффект и базовый алгоритм.

В **четвёртой главе** рассматривается модель показов рекламных объявлений, использующая результаты, полученные в ходе диссертационного исследования. Отдельно рассматривается каждый этап работы модели, а также различные постановки задач оптимизации, которые могут быть поставлены и решены способом, представленным в диссертации. В главе рассматривается комплекс программ,

реализующий подготовку данных и сам алгоритм подбора оптимальных параметров нового вида критерия показа, а также реализация работы критерия отбора с новыми запросами. В конце главы приводятся результаты экспериментального тестирования. Завершается глава описанием результатов внедрения результатов диссертационного исследования на всём потоке запросов поисковой системы «Яндекс».

В **заключении** говорится об основных результатах диссертационного исследования.

ГЛАВА 1. ЗАДАЧА ОТБОРА РЕКЛАМНЫХ ОБЪЯВЛЕНИЙ В РЕКЛАМНЫЙ ПОКАЗ.

1.1 Использование поисковых систем в Internet для рекламных целей. Структура и порядок функционирования рекламного блока (на примере Яндекс.Директ).

В настоящее время огромное количество человек пользуются сетью Интернет для поиска информации [70]. Интернет-пользователь задает запрос в поисковой строке, на который он хочет получить ответ, и задача поисковых систем – предоставлять пользователю наиболее полную и релевантную информацию. Для того чтобы найти эту информацию поисковой системе необходимо обойти огромное количество интернет-сайтов и сохранить те данные, которые на них содержатся. Технология поиска информации в интернете достаточно сложная для реализации и усовершенствования, а для хранения и обработки огромных массивов данных необходимо иметь мощный вычислительный центр и большое количество серверов. Для того чтобы обеспечивать все необходимые составляющие для успешного ответа на запрос пользователей, поисковой системе требуются денежные средства, которые зарабатываются, в основном, посредством показов интернет-рекламы на странице поисковой выдачи.

На странице поисковой выдачи на **запрос** пользователя выделяется специальный рекламный блок (этих блоков несколько, они отличаются по ряду характеристик, речь о которых пойдёт ниже). Единицей для показа в рекламном блоке является рекламное объявление – рекламное сообщение, содержащее информацию о товаре или услуге, предоставляемыми рекламодателем, а также ссылку на соответствующий сайт. Рекламодатели составляют и добавляют в систему множество рекламных объявлений. Рассмотрим более

подробно то, каким образом происходит показ объявления – отображение его на странице результатов поиска компании «Яндекс» ее пользователю. По запросу составляется список **ключевых фраз** (это нормализованные слова из подмножества слов запроса), к которым и «привязываются» рекламные объявления. Рекламодатель может задать для своего объявления ключевую фразу (или несколько ключевых фраз), по которым оно может быть показано. Например, рекламодатель, который хочет охватить тему доставки цветов, может выбрать фразу «цветы», однако эта фраза является слишком широкой для его области деятельности. Таким образом, рекламодателю будет правильнее указать фразу «доставка цветов», а для конкретизации местоположения доставки – «доставка цветов Москва». По списку ключевых фраз составляется множество объявлений-кандидатов для показа в рекламном блоке. Вообще говоря, не все рекламные объявления, отобранные по запросу, могут быть показаны (об этом речь пойдёт позже). Если пользователь кликает на то или иное объявление, происходит переход на рекламируемый сайт, этот переход называется **кликом**. Детальное описание терминологии поисковой контент-рекламы можно найти в [34]. В случае клика осуществляется **списание денежных средств** со счета рекламодателя в пользу поисковой системы, причем общая сумма от списания денежных средств со счетов рекламодателей оказывается существенной составляющей дохода поисковой компании. Рекламодатель для каждого объявления выставляет ставку, тем самым происходит торг за показ его рекламного объявления на странице результатов поиска по запросу.

1.2 Развитие интернет-рекламы, разные схемы списания денежных средств со счёта рекламодателей.

Первой компанией, которая начала показывать интернет-рекламу, была компания GoTo, основанная в 1997 году Биллом Гроссом – известным инвестором интернет-проектов [69]. Билле Гроссу также обычно приписывают саму концепцию продажи ссылок на сторонние сайты, которые как-то связаны с поисковыми запросами. Несмотря на внешнюю простоту идеи, она оказалась инновационной: достаточно уместно показывать рекламу человеку в тот момент, когда он сам проявляет интерес к товару или услуге, задавая запрос поисковой системе. В феврале 1998 года Джеффри Брюер, исполнительный директор GoTo.com представил на конференции TED8 («Технология, развлечения и дизайн») новую концепцию поискового алгоритма. До этого момента поисковые системы выводили результаты согласно сложным механизмам, определяющим релевантность запросу пользователя. GoTo.com предложила новый принцип: выставлять результаты поиска на аукцион. Рекламодателю предлагалось выбрать подходящее слово или фразу, сделать на нее ставку и предоставить соответствующее рекламное объявление. При запросе пользователя по соответствующему слову или фразе результаты выводились бы согласно ставкам рекламодателей, а при клике на ссылку, с них бы списывалась сумма, равная их ставке [22].

Получив $6 миллионов стартового капитала на развитие идеи, GoTo.com начала кампанию по продвижению своего поискового механизма под слоганом «Search made simple» («Сделай поиск легким»). Результаты рекламной кампании были весьма успешны: к концу 1998 года на GoTo.com заходило больше 4 миллионов посетителей в месяц. Однако большинство рекламодателей на рынке жаловались на стремительно снижающуюся эффективность баннерной

рекламы. Средний процент переходов по баннерной рекламе снизился до 0,5 %. Контекстная реклама и принцип оплаты за клики становились все более привлекательными, в особенности для небольших фирм. Вскоре компания подписала соглашение с ведущими провайдерами и поставщиками браузеров об установке поискового механизма от GoTo.com. К 2002 году реклама от компании GoTo.com показывалась на всех крупнейших порталах в Интернете, и в это время свои первые шаги начали делать конкуренты.

Компании, занимающиеся интернет-рекламой, достаточно быстро переняли концепцию платы за клик. BeFirst (теперь MIVA) запустила аналогичный продукт в 1999 году. В 2005 Ask.com так же позаимствовала предложенную GoTo.com схему [71], и MSN Search расширил концепцию для поддержания поведенческого таргетинга [28].

Некоторые компании списывали с рекламодателя фиксированную плату за показ на своём сайте, другие, например, такие как Netscape и Infoseek в 2005 году начали использовать систему $cost - per - mille\ (CPM)$ – стоимость за показ рекламного объявления тысячу раз [40].

Поисковые системы использовали баннерную рекламу до того, как появился отдельный поиск по объявлениям, соотвествующих запросу пользователя. Таким образом, они столкнулись с дилеммой: удерживать пользователя на своём сайте, чтобы он просмотрел как можно больше платных объявлений, или сразу предоставить ему результаты поиска. Поисковые системы справились с этой дилеммой, получая доход от перехода пользователя на сайт рекламодателя (клик по рекламному объявлению) [22]. В 1996 году поисковая система Open Text представила привилегированные списки, в которых были собраны

сайты, которые будут платить за то, чтобы быть добавленными в результаты поиска по определенным ключевым словам [72].

Компания Google начала продавать рекламу в своей поисковой системе еще в 1999 году, однако первое время она следовала модели оплаты за показы рекламных объявлений [11]. Новая же система во многом повторяла концепции Overture, однако привнесла и несколько ключевых нововведений. Так, позиция в рекламной выдаче зависела не только от величины ставки, но и от значения CTR: объявление с большим CTR могло занимать в выдаче место выше, чем объявление с низким CTR рекламодателя, сделавшего большую ставку. Рекламодатели не платили за переход на сайт рекламодателя полный размер ставки: плата за переход составляла ставку предыдущего объявления в выдаче плюс один цент. Рекламодатели могли выбрать страны и языки пользователей, которым нужно было показывать объявление. Был усовершенствован подбор ключевых слов и фраз: стало возможным задавать строгость соответствия ключевых слов запросу пользователя и задавать стоп-слова, при наличии которых в запросе объявление не будет показываться.

Плата за показ рекламного объявления стала не выгодной для рекламодателей, так как с ростом интернет-аудитории количество показов стало достаточно большим, а реальных переходов на сайты это не гарантировало. Базовым вопросом оставалось то, на сколько эффективны показы рекламного объявления и как рекламодателю проверить что при показе его реклама заинтересовала пользователя. Самым простым ответом на этот вопрос является клик по ссылке на сайт, которая указана на рекламном объявлении [74]. В 1996 году компания «Yahoo!» первой начинает списывать с некоторых рекламодателей (например «Procter & Gamble») их ставку только в случае клика по их объявлениям. Новая стратегия списания денег

начала называться $cost - per - click$ (CPC). Первой такую стратегию предложила компания LinkStar в 1996 году [73].

Следующей предложенной схемой для списания денег с рекламодателя была цена за конверсию $cost - per - action$ (CPA) – какое-либо действие на сайте рекламодателя (например покупку товара, занесение товаров в корзину или пребывание на сайте в течении какого-то определённого промежутка времени) [30]. Списание за конверсию выгодно для рекламодателя, однако, оно не так выгодно для поисковой системы, так как сведения, предоставляемые рекламодателями о действиях, происходящих на их сайтах, могут быть недостоверными.

Первой в 1997 году предложила схему списания денег CPA компания DoubleClick [59]. В 2003 году Overture, Google, и FindWhat предложили автоматические инструменты для измерения CPA рекламодателями [48], чтобы те могли выставить соответствующие ставки за клик. SNAP заявил CPA как основную схему списания денег с рекламодателя в своей поисковой системе в 2004 году.

Все вышеперечисленные схемы CPM, CPC и CPA могут быть в разной степени привлекательны для рекламодателей – в зависимости от того, что они хотят получить от рекламной компании. При CPM – это скорее узнаваемость бренда, CPC – эффективность рекламного объявления, CPA – заинтересованности клиента [13]. На сегодняшний день текущая схема списания денег с рекламодателя за клик является наиболее актуальной и оптимальной с точки зрения полезности для поисковой интернет-системы и рекламодателей интернет-рекламы.

1.3 Задача распределения рекламной информации по разным рекламным блокам на странице результатов поиска.

На странице поисковой выдачи выделяется несколько мест, где могут быть показаны рекламные объявления:

Гарантированные показы — места для объявлений, расположенные справа или снизу от результатов поиска.

Спец-размещение — место для показа объявлений, расположенное сверху над результатами поиска (Рис. 1.).

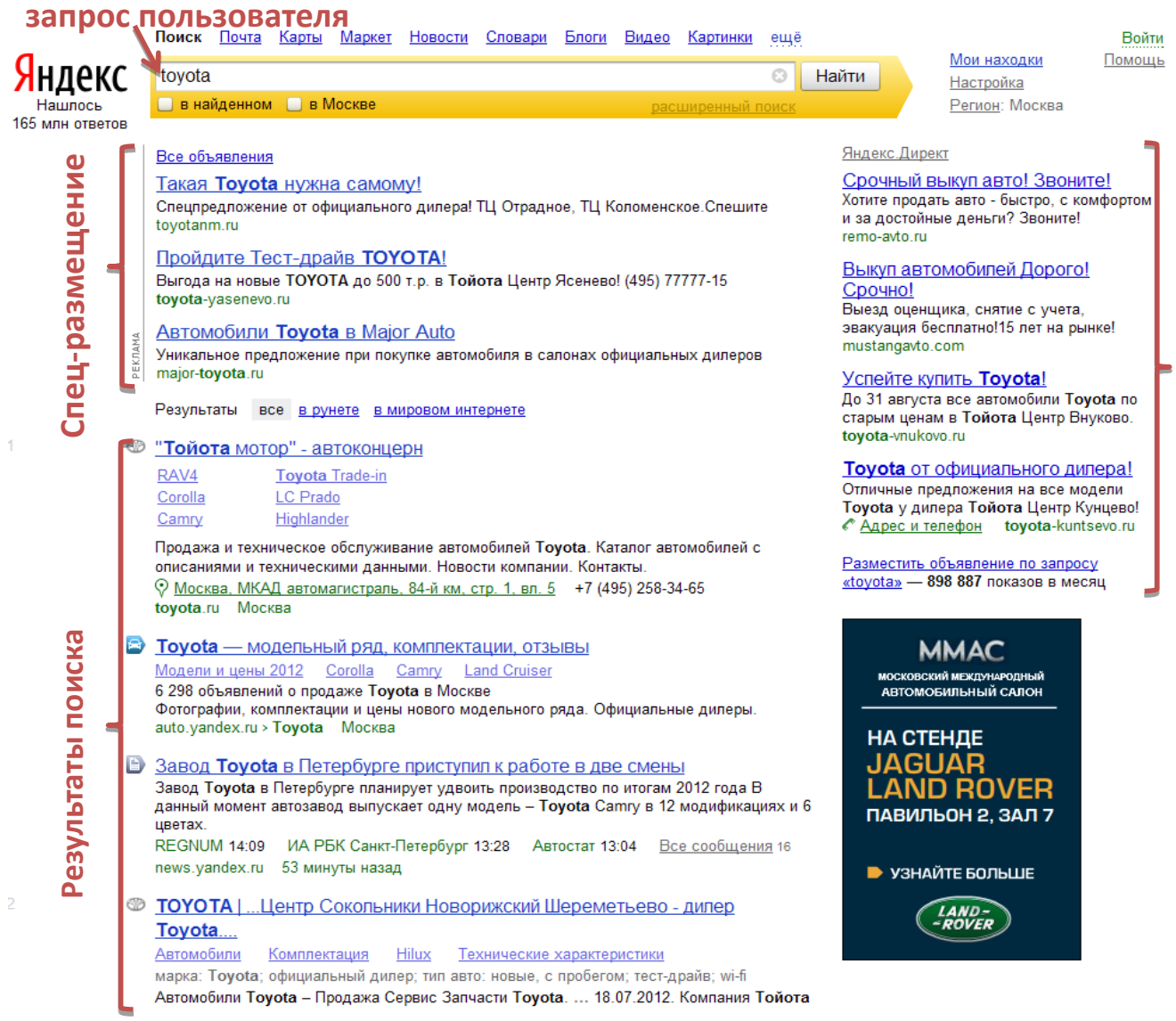

Рис. 1. Место для показов рекламы на странице результатов поиска компании «Яндекс».

Спец-размещение является привилегированным местом для размещения рекламы и приоритетно для рекламодателей [7] вследствие того, что оно расположено над результатами поиска: после того, как пользователь задаёт запрос поисковой системе, первое что он видит – это рекламные объявления, поэтому его внимание будет обращено именно на них. Так как цель рекламодателя – посещение пользователем его сайта, то такой способ размещения наиболее предпочтительный. Среди рекламодателей существует большое количество желающих попасть в блок спец-размещения, поэтому среди них происходит конкуренция за показ. Инструментом конкуренции является выставление ставки, по которой торгуется рекламное объявление. Чем более привлекательным является для рекламодателя показ в блоке спец-размещения – тем более высокую ставку он выставит для своего объявления. Из-за высокой конкуренции и высоких ставок спец-размещения является так же наиболее прибыльным местом показа рекламы для поисковой системы.

Диссертационное исследование посвящено моделированию отбора рекламных объявлений для показа в блоке спец-размещения.

Целесообразность показа рекламного объявления над результатами поиска определяется двумя основными критериями. Первый – это вероятность того, что пользователь кликнет на предъявленное объявление (**кликабельность – $\boldsymbol{CTR, Click-Through-Rate}$**). Можно сказать, что это – степень эффективности показа объявления, его характеристика привлекательности для пользователя. Второй критерий – это **ставка ($\boldsymbol{Bid}$)**, назначенная рекламодателем за клик по его рекламному объявлению, показанному в рекламном блоке над результатами поиска.

Для вероятности клика пользователя по данному объявлению, конечно, можно дать только оценку, поскольку как пользователи, так и

объявления постоянно меняются. На оценку вероятности клика влияет большое количество разнообразных факторов. В настоящее время предсказание CTR – это открытый вопрос для исследователей. На данный момент существует большое количество различных методов и подходов к предсказанию кликабельности, а также к выбору факторов, от которых она зависит [27], [55], [53], [21], [61].

Среди них различаются следующие категории:

- ***История показов***: CTR зависит от истории кликов и показов самого конкретного объявления. При накоплении достаточной статистики можно говорить о том, что наблюдаемый CTR приближается к «истинному» CTR объявления [12], [1].
- ***Текст объявления:*** количество слов в ключевой фразе, по которой было показано объявление; количество слов в заголовке и тексте объявления; доля заглавных символов в тексте объявления; различные морфологические характеристики, например: эмоциональная окраска текста объявления, количество прилагательных, количество глаголов в повелительном наклонении и т.д. [34], [41].
- ***Запрос-ключевое слово:*** мощность пересечения слов запроса и ключевой фразы; количество слов, отличающихся в запросе и ключевой фразе [25].
- ***Запрос-текст объявления*****:** $TF \cdot IDF$ запроса и заголовка/тела объявления, $TF \cdot IDF$ запроса и полного текста объявления и т.д. [39].
- ***Рекламодатель*****:** характеристики рекламной компании, бюджет, стратегия [62].
- ***Пользователь*****:** социально-демографические характеристики (такие как пол, возраст, средний доход), предпочтения пользователя по тематикам, предрасположенность к клику в данный конкретный

момент, кликал ли он на это объявление в прошлом, на сколько часто он кликает на рекламу и т.д. [10], [65], [18].

- ***Запрос:*** количество слов в запросе, количество символов в запросе и т.д.; навигационность, насколько запрос является коммерческим, частотность [15].
- ***Позиционно-временной контекст:*** дата и время показа объявления (существует существенная разница в CTR одного и того же объявления, показанного утром и вечером или в будние дни и выходные). Сюда же можно отнести то, на какой странице поисковой выдачи показано объявление [6].
- ***Сопутствующие элементы:*** сюда входят такие признаки, как характеристики объявлений, показанных вместе с текущим, а так же результаты поисковой выдачи [9].

Всё многообразие признаков используется машинным обучением для построения алгоритма предсказания вероятности клика по объявлению. Однако в данной работе мы не будем касаться этого вопроса, и примем что оценка CTR для каждой пары «ключевая фраза – рекламное объявление» нам задана.

Ожидаемый доход поисковой системы, списываемый со счета рекламодателя за клик пользователя по конкретному рекламному объявлению, может быть оценен как произведение вероятности клика на ставку, назначенную за объявление. Именно эта величина $\boldsymbol{CPM}\ (\boldsymbol{Cost-Per-Million}) = CTR \cdot Bid$ в настоящее время служит критерием, по которому отбираются кандидаты для показа в рекламном блоке над результатами поиска. В действительности списание денежных средств происходит несколько сложнее – согласно аукциону второй цены: списывается ставка предыдущего объявления (следующего по позиции) [29], [47].

Условия показа рекламного объявления в рекламном блоке над результатами поиска определяются двумя ограничениями: во-первых, количество мест для показа рекламных объявлений в рекламном блоке ограничено, поэтому в действующей системе для одного запроса допускается показ не более **трех** рекламных объявлений. Таким образом один показ **рекламного блока** состоит из непустого множества рекламных объявлений, показанных одновременно по одному запросу пользователя. Во-вторых, считается, что не по всем поисковым запросам нужно показывать рекламный блок над результатами поиска, так как объявления не всегда содержат полезную для пользователя информацию и менее релевантны его запросу чем результаты поиска. Поэтому в системе показов рекламных объявлений принято условие, что **покрытие** – доля запросов пользователей, по которым показывается реклама над результатами поиска, ко всем запросам пользователей, должна быть ограничена сверху. Для этого вводится критерий показа рекламного объявления в рекламном блоке над результатами поиска. Если CPM объявления (для данного запроса) превышает некоторый порог, то это объявление становится кандидатом на показ в рекламном блоке над результатами поиска, иначе рекламное объявление заведомо не будет показано. Критерий подбирается таким образом, чтобы часть результатов поиска не сопровождалась рекламой над ними вообще.

Если число кандидатов, прошедших порог, не больше трех, то все они будут показаны. Если же их число превышает три, то в рекламном блоке над результатами поиска показываются только три рекламных объявления с наибольшим значением CPM.

1.4 Существующие постановки задачи оптимизации показов рекламы и методы определения критерия показа в рекламном блоке.

1.4.1 Различные постановки задачи оптимизации показов рекламы.

В литературе можно выделить три подхода к оптимизации показов рекламных объявлений над результатами поиска.

Первый подход использует в качестве способа обучения алгоритма ранжирования рекламных объявлений оптимизацию **релевантности** рекламы [46], [52], [16], цель состоит в максимизации количества кликов. В случае если алгоритм машинного обучения может точно предсказывать клик по рекламному объявлению и его релевантность, то этот подход также приводит к максимизации дохода поисковой системы. Для того чтобы учесть суммарный доход поисковой системы, один подход объединяет в себе доход и релевантность в форме линейной комбинации [52], другой – умножение показателя релевантности на ставку рекламодателя [14]. Оба этих метода являются эвристическими, поэтому нахождение оптимальных параметров достаточно затруднительно. Когда настраиваемых параметров достаточно много, эвристический выбор этих параметров становится невозможным, так как вычислительная сложность увеличивается экспоненциально с ростом числа параметров.

Второй подход разделяет ожидаемый доход на две части, каждую из которых он старается максимизировать: $\boldsymbol{CTR}$ **и ставка рекламодателя**. Предложено несколько алгоритмов для предсказания CTR объявления с высокой точностью [27], [55], [53], [32]. Предположением этого подхода, как указано в [32], является то, что существует собственная релевантность объявления, которая не зависит

от контекста запроса пользователя. Объявление с высоким значением CTR может быть и неревантным конкретному запросу пользователя.

В итоге, механизм, изложенный в первой работе [46], дает результаты только в случае высокой точности предсказания релевантности объявления. Вторая и третья работы [52], [16] используют двух-шаговую оптимизацию дохода вместо релевантности. Несмотря на то, что в этих работах действительно есть определённое увеличение доходности, однако, релевантность объявлений значительно падает.

Третий подход состоит в следующем: как обсуждалось в литературе [43], [35], существует два важных фактора, от которых зависит решение пользователя о том кликнуть на определённое объявление или нет. Во-первых, это – **позиционность** (зависимость вероятности клика от позиции, на котором показывается объявление), которое приводит к увеличению числа кликов на рекламные объявления, размещённые на верхних позициях. Другая состоит в том, что на вероятность клика пользователя влияет не только абсолютная релевантность конкретного рекламного объявления, но и общее **качество других объявлений** в рекламном блоке. Многие авторы строят ранжирующую функцию исходя именно из этих двух факторов и достигают достаточно хороших результатов [38], [43], [49]. В работе Чжу [67] предлагается два подхода к сочетанию прибыли и релевантности рекламных объявлений, связанных с построением сложных целевых функций.

Другие не оптимальные модели рассматриваются в работах [55], [68], [58], [5].

1.4.2 Существующие методы выбора критерия показа в рекламном блоке.

Когда оптимизационная задача поставлена, необходимо зафиксировать ранжирующий критерий отбора объявлений. Классической ранжирующей функцией для отбора объявлений для показа в рекламном блоке над результатами поиска является CPM.

Одна из крупнейших поисковых компаний Yahoo! сортирует результаты поиска в **зависимости от точности соответствия фраз запроса и купленной ключевой фразы**, и только после этого по ставке, соответствующей рекламному объявлению [36]. Система показов интернет-рекламы на поиске Google AdWords в 2002 году изменила критерий ранжирования рекламных объявлений со ставки за клик на **ставку за клик умноженную на кликабельность**. Yahoo! начал вычислять метрику *Click Index* приводя **позиционную наблюдаемую кликабельность к стандартной**. Эта стандартная кликабельность показывается в интерфейсе рекламодателя, как их «истинная» кликабельность, но ранжирование происходит уже по другому критерию.

Таким образом, при сортировке по CPM отбор объявления зависит только от вероятной выгоды его показа для поисковой системы, что является не совсем корректным по отношению к пользователям и рекламодателям.

В научной литературе встречается достаточно большое количество работ с различными критериями отбора объявлений для показа в рекламном блоке над результатами поиска.

Первым из критериев отбора, описанным в статье Агарвала [7] является $\boldsymbol{bid \cdot quality + v}$, где $\boldsymbol{quality}$ – **качество** объявления (в нашем случае это предсказываемый CTR), а $\boldsymbol{v}$ – количественная **мера полезности** объявления для поисковой системы в качестве

обучающего материала и дополнительной информации о качестве (вообще говоря, это – некоторое начальное предсказание CTR для новых рекламных объявлений). При выборе v нужно следить за балансом между показами новых и старых объявлений. В продолжение этого критерия существует такая схема отбора объявлений, как по $\boldsymbol{bid} \cdot \boldsymbol{quality} + \boldsymbol{v}$, где $\boldsymbol{v} = \boldsymbol{c} \cdot \boldsymbol{bid} \cdot \boldsymbol{var}(\boldsymbol{quality})$, а c – константа, показывающая на сколько полезно обучение на данном объявлении, то есть **баланс между долгосрочным и краткосрочным показом объявления**. Данный метод интересен тем, что начальный прогноз качества объявления не является одной константой, а зависит от дисперсии этого прогноза и его ставки, то есть отбирается то объявление на показ, которое принесёт поисковой системе больший доход.

Вторым направлением для выбора критерия показа объявлений является **релевантность** – степень соответствия содержания рекламного объявления запросу пользователя (семантическое соответствие поискового запроса и текста объявления) [14], [21], [46]. Однако отбор только посредством релевантности учитывает только интересы пользователя. Рекламодатель практически не может напрямую повлиять на отбор его объявления на показ, так как его ставка не входит в критерий показа рекламных объявлений над результатами поиска. Поисковая интернет-система тоже имеет достаточно ограниченное влияние на свои основные показатели, такие как суммарные клики, среднюю кликабельность и доход: отбор объявлений зависит только от запроса пользователя, и от того насколько корректно составлен текст объявления и оптимизирован сайт рекламодателя под текущий запрос.

Третьим подходом является выбор в качестве характеристик, от которых зависит критерий показа объявлений, следующих: на сколько

фраза, по которой показывается объявление, является **широко-тематической** или **узко-тематической**. Тут возникает проблема баланса между фразами: зачастую бывает достаточно сложно определить тематичность фраз, для этого нужно собирать и хранить большое количество информации, а также обновлять её по мере накопления статистики. Следующей характеристикой рекламных объявлений является **информация о продавце** и информация о **бренде**, представляемом объявлением [33]. Однако тут есть сложность в достоверности данных для разных пользователей, а также изменение предпочтений и вкусов. Не всегда представляется возможным собрать достаточную статистику по всем рекламодателям и брендам, а для тех, кто впервые встретился пользователю, информации нет никакой и как их ранжировать остаётся под вопросом.

Четвёртым подходом среди работ, посвящённых отбору объявлений в рекламный блок, является применение **позиционных моделей** показа рекламы [66]. Чжаном было выявлено, что позиция, на которой было показано объявление, влияет на его кликабельность, следовательно, важно её учитывать при отборе объявлений. Также для разных рекламных объявлений показ на той или иной позиции может иметь различный эффект, важно учесть совместный показ группы объявлений, причём каждого на своей собственной позиции. Позиционные эффекты будут рассмотрены в данной работе как расширение одного из критериев показа рекламных объявлений в блоке над результатами поиска.

Пятым подходом, совместно с позиционной моделью, часто рассматривается модель, зависящая от **пользователя**, то есть в критерий показа добавляется параметр, зависящий от характеристик пользователя. Например в [57] критерий показа имеет вид: $CTR_i \cdot pos_i \cdot ucp \cdot bid_i \geq \theta_i$, где добавляется мультипликативный множитель ucp –

персональный признак, обозначающий на сколько данный пользователь расположен кликать на рекламу в данный момент по сравнению со среднестатистическим пользователем.

Шестым в качестве критерия показа рекламного объявления над результатами поиска пробовался критерий вида $\boldsymbol{bid^k \cdot CTR^l}$. Данный вид критерия даёт достаточно большую степень свободы при отборе объявлений. Однако ограниченность данного метода заключается в том, что ставка и предсказанная вероятность клика входят в ранжирующий критерий только как произведение одно на другое. То есть при любых k и l мы можем усиливать зависимость выбора либо от ставки данного объявления, либо от предсказанного CTR объявления. Более подробно про баланс между ставкой и качеством выдачи (в нашем случае CTR) можно почитать в работе Агарвала [7].

Как видно из представленного обзора, тема критерия показа рекламных объявлений над результатами поиска актуальна и широко исследуется научным сообществом. Также важно заметить, что некоторые авторы ставят оптимизационные задачи, однако в рассмотренных работах есть некоторые ограничения и алгоритмы больше эвристические и подходят только лишь к конкретным постановкам задачи оптимизации.

1.5 Постановка задачи исследования.

Отбор объявлений для показа в рекламном блоке над результатами поиска производится посредством критерия показа. В современных исследованиях не стоит конкретной оптимизационной задачи: каждый из исследователей предлагает вид критерия показа исходя из своих предположений и предпочтений. В настоящем диссертационном исследовании ставится задача формального описания оптимизационной задачи показа рекламных объявлений, её

математической формулировки, а также решения исходя из внешних ограничений по некоторым характеристикам системы показов рекламы.

Выше было приведено некоторое множество типов оптимизационных задач для поисковых интернет-систем. Теперь необходимо зафиксировать какая именно оптимизационная задача будет нами рассмотрена. Положим, что целесообразность показа рекламных объявлений в рекламном блоке над результатами поиска задается несколькими критериями:

- **Суммарные денежные средства**, списываемые со счетов рекламодателей (суммарный доход поисковой интернет-системы от показов рекламы на странице своей поисковой выдачи). Тогда оптимизация правила отбора рекламных объявлений для показов над результатами поиска представляется несложной задачей: нужно просто подобрать такое значение критерия показа, при котором покрытие будет в точности равно заданному ограничению (то есть доля запросов с рекламой над результатами поиска фиксирована).
- **Средняя кликабельность** – то, насколько часто в среднем пользователь кликает на одно из показанных объявлений. Этот показатель определяет удовлетворённость пользователя, а также эффективность самой рекламы. Средняя кликабельность соответствует среднему значению CTR по всем запросам и всем объявлениям, попавшим в рекламный блок над результатами поиска. Если эта величина мала, то пользователь либо перестанет кликать по рекламе, либо будет искать другие способы поиска информации в интернете. Ни то ни другое не удовлетворяет потребностям поисковой системы.

В данной работе мы ставим задачу **максимизации среднего значения $\boldsymbol{CTR}$**, но также действует ограничение на суммарный доход

поисковой системы, поэтому задача ставится как поиск условного максимума средней кликабельности в рекламном блоке над результатами поиска при условии, что сумма денежных средств должна быть не меньше заданной величины.

Кроме того, сохраняются **ограничения на покрытие** и на **число рекламных объявлений** в одном рекламном блоке над результатом поиска. Задача состоит в том, чтобы найти критерий показа, доставляющее максимум среднему значению CTR при упомянутых ограничениях.

ГЛАВА 2. ПОСТАНОВКА ЗАДАЧИ ОПТИМИЗАЦИИ ПОКАЗОВ РЕКЛАМНЫХ ОБЪЯВЛЕНИЙ И ПОСТРОЕНИЕ АЛГОРИТМА ПОДБОРА ПАРАМЕТРОВ КРИТЕРИЯ ПОКАЗА.

2.1 Математическая постановка задачи.

2.1.1 Обозначения.

Поставим задачу следующим образом: как для заданной выборки запросов и заданного списка рекламных объявлений найти **оптимальную расстановку** этих объявлений в рекламном блоке над результатами поиска для показа по этим заданным запросам [2], [19].

Считаем, что **ставки ($\boldsymbol{Bid}$)** для всех рекламных объявлений известны, а также известны оценки **вероятности клика ($\boldsymbol{CTR}$)** для каждой пары «запрос-объявление» (если данное рекламное объявление в принципе не может быть совмещено с определенным ответом на запрос, то полагаем, что оценка CTR равна нулю). Считаем также, что заданы: минимальная сумма денежных средств, которая должна поступить от рекламодателей, покрытие – доля результатов поиска, сопровождаемых рекламой над ними, а также максимальное число рекламных объявлений (k) в каждом рекламном блоке. При этих ограничениях будем искать такую расстановку рекламных объявлений в блоке над результатами поиска, которая обеспечивает максимум средней кликабельности – среднего значения CTR.

Задачу будем решать в том приближении, когда рекламодатель платит за клик по своей ставке, а не по правилу аукциона второй цены (сами деньги будут слегка завышены, но разницу в характеристиках мы можем достаточно точно проследить).

Полагаем, что в нашей выборке имеется $\boldsymbol{Q}$ запросов, которые мы будем нумеровать индексом q ($1 \leq q \leq \boldsymbol{Q}$), и задан список из $\boldsymbol{A}$ баннеров, нумеруемых индексом a ($1 \leq a \leq \boldsymbol{A}$) Обозначим далее:

$\boldsymbol{Bid_a}$ – ставка, назначенная рекламодателем за a–ое рекламное объявление. Все ставки неотрицательны, т.е. $Bid_a \geq 0$.

$\boldsymbol{CTR_{aq}}$ – оценка вероятности клика при размещении a–ого рекламного объявления над результатами поиска на q–ый запрос. Величины Bid_a и CTR_{aq} считаем заданными.

Введем далее бинарную переменную $\boldsymbol{t_{aq}}$, такую что $\boldsymbol{t_{aq}} = 1$ означает, что a–ое рекламное объявлений размещено над результатом поиска на q–ый запрос, и $\boldsymbol{t_{aq}} = 0$ в противном случае. Именно оптимальные значения t_{aq} ищутся в диссертационной работе, обозначим все искомые переменные t_{aq} как матрицу $\boldsymbol{T}$.

Тогда:

$$\boldsymbol{Events} = \sum_{(a,q)} t_{aq}$$

будет означать суммарное количество показов рекламных объявлений по данной выборке запросов в рекламном блоке над результатами поиска, а

$$\boldsymbol{CRIT_0(T)} = \sum_{(a,q)} \boldsymbol{CTR_{aq} \cdot t_{aq} / Events} \qquad (1)$$

– среднюю кликабельность (среднее значение по CTR по этим рекламным объявлениям). Именно эту величину мы хотим максимизировать при заданных ограничениях.

2.1.2 Ограничения.

Ограничение по суммарному доходу. Математическое ожидание денежных средств, списываемых со счета рекламодателя, равно вероятности клика на данное рекламное объявление по заданному запросу, умноженной на ставку рекламодателя выставленную для этого объявления, в случае его показа в рекламном блоке над результатами поиска. Если же оно не показано, то вероятность клика по объявлению

равна нулю, а, поэтому, и денежных средств оно принести не может. За оценку вероятности клика мы принимаем заданное значение CTR. Таким образом величина денежных средств, которые поисковая система может получить от показа рекламного объявления, принимает значение $Bid_a \cdot CTR_{aq} \cdot t_{aq}$, где a – номер баннера, а q – номер запроса из нашей выборки. Тогда общая сумма денежных средств, списываемых с рекламодателей по нашей выборке запросов M будет равна:

$$M(\boldsymbol{T}) = \sum_{(a,q)} Bid_a \cdot CTR_{aq} \cdot t_{aq}$$

требуется, чтобы эта сумма была не меньше заданного значения M_{min}. Таким образом, ограничение принимает вид:

$$\boldsymbol{M(T)} \ \geq \ \boldsymbol{M_{min}} \qquad (2)$$

Ограничение на покрытие. Напомним, что покрытием называется доля запросов, по которым в рекламном блоке над результатами поиска показано хотя бы одно рекламное объявление среди всех результатов поиска, по которым была показана реклама.

Выражение $\mathbb{I}\{\sum_{(a)} t_{aq} > 0\} = 1$, если над результатом поиска размещен хоть один баннер, и 0 в противном случае. Тогда ограничение на покрытие $\boldsymbol{H}$ примет вид:

$$\boldsymbol{H(T)} \ = \ \sum_{(q)}\left(\mathbb{I}\left\{\sum_{(a)} \boldsymbol{t_{aq}} > \boldsymbol{0}\right\}\right) \ \leq \ \boldsymbol{H_{max}} \qquad (3)$$

Ограничение на число рекламных объявлений, размещаемых в рекламном блоке над каждым результатом поиска. Считается, что это число не должно превышать k. Тогда это ограничение запишется как:

$$\forall \, \boldsymbol{q} \ \sum_{(a)} \boldsymbol{t_{aq}} \ \leq \ \boldsymbol{k} \qquad (4)$$

В поисковых системах количество рекламных объявлений, размещаемых над результатами поиска, исторически было ограничено

[54], причём максимально допустимым считается четыре рекламных объявления над поисковой выдачей. Мы же ограничимся $k = 3$ (этот параметр задаётся критериями релевантности поисковой выдачи по сравнению с рекламной выдачей).

2.1.2 Математическая модель показов рекламных объявлений.

Теперь, когда приведены все математические обозначения, можно описать математическую модель показов рекламных объявлений в рекламном блоке над результатами поиска.

Входные данные. Имеется набор случайных запросов из логов запросов пользователей q $(1 \leq q \leq Q)$. Для каждого из запросов составляется список рекламных объявлений a $(1 \leq a \leq A)$ для которых известно: предсказание вероятности клика CTR_{aq} и ставка Bid_a назначенная рекламодателем за попадание в рекламный блок над результатами поиска.

Процесс моделирования. Теперь, когда имеется набор запросов и набор рекламных объявлений, необходимо решить какие конкретно из объявлений следует показать на запросы. Существует некоторый **критерий показа**. В базовой модели показов рекламы он одинаковый для всех рекламных объявлений и запросов, то есть $Bid_a \cdot CTR_{aq} > \theta$. Если это условие выполнено, то индикатор показа t_{aq} равен 1 и 0 иначе. Таким образом заполняется матрица $\boldsymbol{T}$ показов по всему набору запросов Q и множеству рекламных объявлений A.

Выходные характеристики системы. После того как матрица T определена и известно какие объявления покажутся по каждому из запросов, можно посчитать суммарные характеристики системы: суммарный доход $M(\boldsymbol{T}) = \sum_{(a,q)} Bid_a \cdot CTR_{aq} \cdot t_{aq}$, среднюю кликабельность по всему набору запросов $CTR_{avg}(\boldsymbol{T}) =$

$\sum_{(a,q)} CTR_{aq} \cdot t_{aq}/Events$ и количество запросов с показанным по ним рекламным блоком $H(\boldsymbol{T})$ (3).

Недостатком данной математической модели показов рекламы (Рис. 2.) является то, что параметр, с помощью которого можно управлять основными характеристиками всей системы, только один. С помощью всего одного параметра не представляется возможным выбирать объявления для показа на запрос таким образом, чтобы оптимизировать какую-либо из характеристик системы.

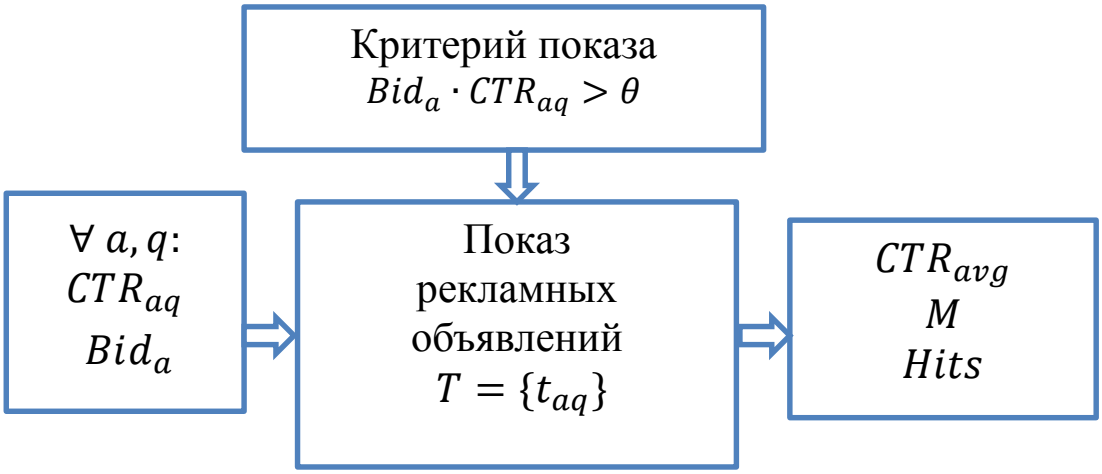

Рис. 2. Математическая модель показов рекламных объявлений.

2.1.3 Формальное описание задачи оптимизации.

Чтобы получить новую математическую модель показов рекламы поставим следующую задачу оптимизации: максимизировать по бинарным переменным t_{aq} функцию

$$\boldsymbol{CRIT_0(T)} \;=\; \sum_{(a,q)} \boldsymbol{CTR_{aq} \cdot t_{aq}/Events} \qquad (5)$$

при ограничениях:

$$\sum_{(a,q)} Bid_a \cdot CTR_{aq} \cdot t_{aq} \geq M_{min}$$

$$\sum_{(q)} \mathbb{I}\left\{\sum_{(a)} t_{aq} > 0\right\} \leq H_{max}$$

$$\forall\, q \;\; \sum_{(a)} \boldsymbol{t_{aq}} \;\leq\; k$$

Это задача дискретного программирования.

2.2 Решение задачи оптимизации. Алгоритм подбора оптимальных параметров.

2.2.1 Переход от дискретной задачи к непрерывной.

Сформулированная постановка задачи относится к классу задач дискретной оптимизации. Решать такие задачи трудно, обычно их решение требует комбинаторного перебора. Но, если такую задачу погрузить в непрерывное пространство, то часто для ее решения удается воспользоваться хорошо разработанными методами оптимизации в непрерывном пространстве искомых переменных. Мы также воспользуемся этим приемом и заменим бинарные переменные t_{aq} на непрерывные t_{aq}, подчиненные дополнительному ограничению: $0 \leq t_{aq} \leq 1$. Эта замена корректна, потому, что, как мы увидим, при максимизации нашего критерия переменные все равно принимают крайние значения 0 или 1. Поэтому решение непрерывной задачи совпадет с решением дискретной [3], [4].

2.2.2 Общий принцип – метод множителей Лагранжа.

Согласно методу множителей Лагранжа, если требуется найти максимум некоторой функции $F(\boldsymbol{x})$ при ограничениях $\varphi_i(\boldsymbol{x}) = 0$, то можно эту задачу заменить на задачу поиска безусловного максимума функции $F^*(\boldsymbol{x}) = F(\boldsymbol{x}) - \sum \lambda_i \varphi_i(\boldsymbol{x})$, причем коэффициенты λ_i подбираются так, чтобы в точке максимума $F^*(\boldsymbol{x})$ все ограничения $\varphi_i(\boldsymbol{x}) = 0$ выполнились точно.

Если же ограничения имеют вид неравенств $\varphi_i(\boldsymbol{x}) \leq 0$, то согласно теории Куна-Таккера [3], метод множителей Лагранжа модифицируется следующим образом: по-прежнему, ищется

безусловный максимум функции $F^*(\boldsymbol{x}) = F(\boldsymbol{x}) - \sum \lambda_i \varphi_i(\boldsymbol{x})$, а коэффициенты подбираются так, чтобы выполнились три условия:

- точка максимума функции $F^*(\boldsymbol{x})$ должна удовлетворять ограничениям $\varphi_i(\boldsymbol{x}) \leq 0$,

- коэффициенты должны быть неотрицательны ($\lambda_i \geq 0$),

- коэффициент λ_i должен быть равен нулю, если в точке максимума $F^*(\boldsymbol{x})$ соответствующее ограничение не достигает предельного значения, т.е. $\varphi_i(\boldsymbol{x}) < 0$.

На самом деле можно не все ограничения заменять слагаемыми в критерии, а часть из них оставить как ограничения [4]. То есть искать условный максимум функции

$$F^*(\boldsymbol{x}) = F(\boldsymbol{x}) - \sum_{(i=1,m)} \lambda_i \varphi_i(\boldsymbol{x}),$$

при ограничениях $\varphi_i(\boldsymbol{x}) \leq 0$ $(i = m+1, n)$.

Требования к подбору коэффициентов остаются теми же, что перечислены выше.

Именно этим приемом, когда часть ограничений переводится в критерий (с последующим удовлетворением ограничений путем выбора соответствующих множителей Лагранжа), а часть сохраняется, мы и будем пользоваться в дальнейшем.

2.2.3 Применение метода множителей Лагранжа к задаче оптимизации.

Добавление в критерий ограничения по деньгам.

Итак, требуется найти максимум по $\boldsymbol{T}$ функции $CRIT_0(\boldsymbol{T}) = \sum CTR_{aq} \cdot t_{aq}/Events$ при ограничениях (2), (3), (4) и $0 \leq \boldsymbol{t}_{aq} \leq 1$. Сначала переведем в критерий (с помощью множителя Лагранжа $\lambda_1 \geq 0$) только ограничение (2). Запишем его в виде $\varphi(\boldsymbol{x}) \leq 0$:

$$M_{min} - M(\boldsymbol{T}) \leq 0,$$

Получим новый критерий

$$CRIT_1(\boldsymbol{T}) = \frac{\sum CTR_{aq} \cdot t_{aq}}{\sum t_{aq}} - \lambda_1\big(M_{min} - M(\boldsymbol{T})\big),$$

и будем искать его максимум с учетом остальных ограничений.

Дополнительное ограничение.

Если бы наш критерий представлял собой сумму функций, каждая из которых зависит только от одной переменной t_{aq}, то можно было бы оптимизировать каждую из этих функций по отдельности – провести декомпозицию задачи. Но в нашем случае это не так, поскольку знаменатель $Events = \sum_{(a,q)} t_{aq}$ в слагаемых $\frac{\sum CTR_{aq} \cdot t_{aq}}{\sum t_{aq}}$ зависит сразу от всех переменных. Для того чтобы все же добиться декомпозиции, предлагается следующее:

а) Сначала зафиксировать знаменатель $Events$ дополнительным требованием:

$$Events = E_0,$$

где E_0 – некоторая назначенная положительная константа, и найти максимум $CRIT_1(\boldsymbol{T})$ с учетом этого дополнительного ограничения. Этот максимум будет зависеть от назначенного значения E_0.

Теперь наш критерий приобретает вид:

$$CRIT_1(\boldsymbol{T}) = \frac{\sum CTR_{aq} \cdot t_{aq}}{E_0} - \lambda_1 \left(M_{min} - \sum_{(a,q)} Bid_a \cdot CTR_{aq} \cdot t_{aq} \right),$$

и действительно становится суммой функций, зависящих только от одной переменной t_{aq}.

Решая задачу, найдем значения переменных t_{aq}, доставляющие условный максимум критерия $CRIT_1(\boldsymbol{T})$ при заданных λ_1 и E_0.

б) Перебором по E_0 найти то значение $E_{0(\text{опт})}$, при котором достигается максимум критерия $CRIT_1(\boldsymbol{T})$. Переведем с помощью

множителя Лагранжа и новое дополнительное ограничение в критерий. Получим:

$$CRIT_2(\boldsymbol{T}) = \frac{\sum CTR_{aq} \cdot t_{aq}}{E_0} - \lambda_1 \cdot \left(M_{min} - \sum_{(a,q)} Bid_a \cdot CTR_{aq} \cdot t_{aq} \right) -$$

$$-\lambda_2 \cdot \left(\sum_{(a,q)} t_{aq} - E_0 \right)$$

при ограничениях:

$$0 \le t_{aq} \le 1,$$

$$\forall q \sum_a t_{aq} \le k$$

$$H(\boldsymbol{T}) = \sum\nolimits_{(q)} \mathbb{I}\left\{ \sum\nolimits_{(a)} t_{aq} > 0 \right\} \le H_{max}$$

Оптимизация критерия $CRIT_2(\boldsymbol{T})$ при оставшихся ограничениях.

Итак, на данном шаге мы должны максимизировать критерий $CRIT_2(\boldsymbol{T})$ считая величины λ_1, λ_2 и E_0 фиксированными. Объединяя члены суммы, зависящие от t_{aq}, и вынося за скобки t_{aq}, получим:

$$CRIT_2(\boldsymbol{T}) = \sum_{(a,q)} t_{aq} \cdot (CTR_{aq}/E_0 + \lambda_1 \cdot Bid_a \cdot CTR_{aq} - \lambda_2) + const$$

где величина $const$ от переменных t_{aq} не зависит. Видим, что наш критерий действительно представляет собой сумму функций, каждая из которых линейно зависит только от одной переменной t_{aq}.

Шаг 1.

Теперь найдем максимум критерия $CRIT_2(\boldsymbol{T})$ по переменным t_{aq} с учетом только одного **ограничения**: $0 \le t_{aq} \le 1$.

Ясно, что при этом переменная t_{aq} должна принять значение 0, если коэффициент при ней отрицателен. Действительно, в этом случае

вклад в критерий $CRIT_2(\boldsymbol{T})$ от члена суммы $t_{aq} \cdot (CTR_{aq}/E_0 + \lambda_1 \cdot Bid_a \cdot CTR_{aq} - \lambda_2)$ будет отрицательным при любом положительном значении t_{aq}, а при $t_{aq} = 0$ этот вклад будет нулевым.

Если же этот коэффициент положителен, то без учета остальных ограничений переменная t_{aq} должна принять значение 1. Действительно в этом случае вклад в критерий будет максимальным (и положительным) при $t_{aq} = 1$.

В случае же, когда этот коэффициент в точности равен нулю при некоторых значениях пары (a, q), мы получаем, что $CRIT_2(\boldsymbol{T})$ не зависит от t_{aq}. При естественных значениях входных параметров, точное равенство коэффициента нулю может случиться только в вырожденных случаях. Но при подборе значений коэффициентов Лагранжа может оказаться необходимым выбрать дробное значение t_{aq}. Однако на практике такое может случиться только для очень малого числа пар (a, q). Действительно, мы подбираем совсем небольшое число неопределенных коэффициентов, и обеспечить точное равенство коэффициента нулю можно тоже в очень малом количестве случаев. Учитывая, что в нашем пуле содержатся тысячи элементов, замена значения t_{aq} на 0 или 1 в этих случаях практически не влияет на значение критерия.

Но остальные ограничения могут сделать показ a–ого объявления на q–тый запрос все же недопустимым, т.е. придется положить $t_{aq} = 0$.

Обозначим $F_{aq}(\lambda_1, \lambda_2, E_0)$ коэффициент при t_{aq}:

$$F_{aq}(\lambda_1, \lambda_2, E_0) = CTR_{aq}/E_0 + \lambda_1 \cdot Bid_a \cdot CTR_{aq} - \lambda_2$$

Тогда кандидатами на показ в рекламном блоке над результатами поиска будут те объявления, для которых $F_{aq} > 0$. Таким образом $F_{aq}(\lambda_1, \lambda_2, E_0)$ является критерием показа рекламного объявления.

Составим теперь для каждого (q–ого) запроса список кандидатов на показ в рекламном блоке над результатами поиска, упорядочив их в порядке убывания величин $F_{aq}(\lambda_1, \lambda_2)$. Некоторые из списков могут оказаться пустыми, если для всех a окажется, что $F_{aq}(\lambda_1, \lambda_2) \leq 0$.

Шаг 2.

Теперь учтем ограничение на **количество баннеров**, размещаемых сверху от результатов поиска:

$$\forall q \sum_{(a)} t_{aq} \leq k$$

Если количество рекламных объявлений в списке кандидатов для q–ого запроса меньше k, то это ограничение выполнится автоматически. Иначе, для максимизации нашего критерия $CRIT_2(\boldsymbol{T})$ следует оставить в каждом списке ровно k кандидатов с максимальным значением $F_{aq}(\lambda_1, \lambda_2)$, т.е. первые k элементов списка. Для остальных элементов положить $t_{aq} = 0$. Назовем этот укороченный список «список для показа».

Шаг 3.

Остается **ограничение на покрытие**:

$$H(\boldsymbol{T}) = \sum\nolimits_{(q)} \mathbb{I}\left\{\sum\nolimits_{(a)} t_{aq} > 0\right\} \leq H_{max}$$

то есть реклама над результатами поиска должна присутствовать не более чем в H_{max} результатах поиска. Если и так количество запросов с не пустым усеченным списком меньше чем H_{max}, то ограничение выполнится автоматически. В противном случае часть списков следует обнулить.

Обозначим $R_q(\lambda_1, \lambda_2)$ вклад в критерий для каждого запроса от рекламных объявлений, вошедших в усеченный список:

$$R_q(\lambda_1, \lambda_2) = \sum\nolimits_{(i)} (CTR_{aq}/E_0 + \lambda_1 \cdot Bid_a \cdot CTR_{aq} - \lambda_2),$$

где сумма для каждого запроса берется только по рекламным объявлениям, вошедшим в список для показа.

Ясно, что максимум критерия будет достигнут, если мы оставим только H_{max} списков с наибольшим значением $R_q(\lambda_1, \lambda_2)$. Иными словами, если мы упорядочим списки в порядке убывания $R_q(\lambda_1, \lambda_2)$, то следует оставить только H_{max} первых списков (если, конечно их хватит, иначе оставляем все). **Обозначим** $\lambda_{3\ \text{опт}}$ минимальное значение $R_q(\lambda_1, \lambda_2)$ по всем оставшимся спискам. Тогда будут оставлены только те списки, для которых $R_q(\lambda_1, \lambda_2) \geq \lambda_3$. Разумеется, величина λ_3 будет зависеть от принятых значений λ_1, λ_2 и E_0.

Теперь можно зафиксировать оптимальные значения t_{aq}:

$$\begin{cases} t_{aq} = 1, & \text{если } a\text{– е объявление в списке для показа для } q\text{– ого запроса} \\ t_{aq} = 0, & \text{в противном случае} \end{cases}$$

Этот результат мы получили для заданных значений λ_1, λ_2 и E_0. Обозначим через $t_{\text{опт}}(\lambda_1, \lambda_2, E_0)$ вектор, составленный из значений t_{aq} этого решения.

Выбор необходимого значения λ_2.

Теперь мы должны найти то значение коэффициента λ_2, при котором действительно выполнится наше дополнительное ограничение $Events = E_0$.

Заметим, что, зная значения t_{aq}, мы можем подсчитать величину $Events = \sum_{(a,q)} t_{aq}$, причем значение $Events$ (как и значения t_{aq}) будет зависеть от зафиксированных значений λ_1, λ_2 и E_0. Но нас сейчас будет интересовать только зависимость $Events$ от Лагранжева коэффициента λ_2, считая по–прежнему значения λ_1 и E_0 фиксированными. Перебирая все значения λ_2, мы можем попытаться добиться равенства $Events(\lambda_2) = E_0$. Конечно, может случиться, что ни при каком значении λ_2 это равенство не выполнится (в силу цело-численности E_0

или по иной причине, но такие значения объявим не допустимыми). Найденное значение $\lambda_{2\ \text{опт}}(\lambda_1, E_0)$ (абсолютного порога) будет зависеть от λ_1 и E_0. Ему будет соответствовать решение $t_{\text{опт}}(\lambda_1, \lambda_{2\ \text{опт}}(\lambda_1, E_0), E_0)$.

Выбор оптимального значения E_0 путем максимизации критерия $CRIT_1(T)$.

Теперь нам нужно избавиться от дополнительного ограничения $Events = E_0$, поскольку оно не входит в число ограничений исходной постановки задачи. Сделать это можно с помощью максимизации критерия:

$$CRIT_1(\boldsymbol{T}, \lambda_1, \lambda_2, E_0) = \frac{\sum CTR_{aq} \cdot t_{aq}}{E_0} - \lambda_1\big(M_{min} - M(\boldsymbol{T})\big)$$

путем перебора по всем допустимым значениям E_0 при фиксированном значении λ_1. При этом в качестве t_{aq} должны браться значения, найденные при $\lambda_2 = \lambda_{2\ \text{опт}}(\lambda_1, E_0)$. Это значение заведомо обеспечивает равенство $Events(\lambda_1, \lambda_2, E_0) = E_0$ в силу операции, описанной в предыдущем разделе.

Обозначим $E_{0\ \text{опт}}(\lambda_1)$ то значение E_0, при котором достигается максимум критерия $CRIT_1$. Оно будет зависеть от значения λ_1. Ему будут соответствовать значения $\lambda_{2\ \text{опт}}(\lambda_1, E_0)$, λ_3 и решение $t_{\text{опт}}(\lambda_1, \lambda_{2\ \text{опт}}(\lambda_1, E_{0\ \text{опт}}(\lambda_1)), E_{0\ \text{опт}}(\lambda_1))$.

Оптимизация критерия $CRIT_0(T)$: выбор значения λ_1.

Найденные значения $E_{0\ \text{опт}}$, $\lambda_{2\ \text{опт}}$, $\lambda_{3\ \text{опт}}$ и $t_{\text{опт}}$ будут зависеть от параметра λ_1. Согласно теории Куна-Таккера, для максимизации нашего основного критерия $CRIT_0(\boldsymbol{T})$ остается выбрать правильное значение параметра $\lambda_1 \geq 0$. Этот параметр нужно выбрать так, чтобы выполнилось ограничение $M(\boldsymbol{T}) \geq M_{min}$. При том этот параметр может быть отличен от нуля только если неравенство по суммарным денежным средствам переходит в равенство:

$$\boldsymbol{M(T) = M_{min}} \quad (6)$$

При $\lambda_1 = 0$ в критерии $CRIT_1(\boldsymbol{t})$ денежные средства вообще не учитываются, и если при этом вырученных денежных средств хватает, то неравенство по деньгам выполнится автоматически. Если же денежных средств не хватает, то следует увеличивать λ_1, пока не выполнится равенство (6). Вырученные денежные средства с ростом λ_1, во всяком случае, не убывают, так как величина $M(\boldsymbol{T})$ входит в критерий $CRIT_1(\boldsymbol{T})$ все с большим весом. Если же при сколь угодно большом значении λ_1 денежных средств все равно не хватает, то задача вообще неразрешима. В нормальном случае перебором по λ_1 найдем то значение $\lambda_{1\,\text{опт}}$, при котором выполнится равенство (6). Соответственно определятся значения $E_{0\,\text{опт}}(\lambda_{1\,\text{опт}})$, $\lambda_{2\,\text{опт}}(\lambda_{1\,\text{опт}}, E_{0\,\text{опт}})$ и решение $t_{\text{опт}}(\lambda_{1\,\text{опт}}, \lambda_{2\,\text{опт}}, E_{0\,\text{опт}})$, и задача будет решена полностью.

Заметим, что если уже выбраны параметры $\lambda_{1\,\text{опт}}$, $E_{0\,\text{опт}}$, $\lambda_{2\,\text{опт}}$, и $\lambda_{3\,\text{опт}}$, то для заданного (q–го) запроса определить, какие рекламные объявления должны показаться над результатами поиска по этому запросу (определить $t_{aq\,\text{опт}}$), можно не обращая внимания на другие запросы. Для этого нужно:

1) Вычислить значение

$$F_{aq} = CTR_{aq}/E_{0\,\text{опт}} + \lambda_{1\,\text{опт}} \cdot Bid_a \cdot CTR_{aq} - \lambda_{2\,\text{опт}}$$

для всех рекламных объявлений (a–ых), совместимых с данным запросом.

2) Составить список кандидатов для показа. В этот список поместить только те рекламные объявления, для которых

$$CTR_{aq}/E_{0\,\text{опт}} + \lambda_{1\,\text{опт}} \cdot Bid_a \cdot CTR_{aq} > \lambda_{2\,\text{опт}}.$$

Этот список может быть и пустым.

3) Если число кандидатов больше k, то сократить список, оставив в нем только k объявлений с наибольшим значением F_{aq}.

4) Подсчитать значение

$$R_q(\lambda_{1\ \text{опт}}, \lambda_{2\ \text{опт}}) = \sum_{(a)} (CTR_{aq}/E_{0\ \text{опт}} + \lambda_{1\ \text{опт}} \cdot Bid_a \cdot CTR_{aq} - \lambda_{2\ \text{опт}}),$$

где сумма берется по рекламным объявлениям из списка кандидатов.

5) Если эта величина меньше $\lambda_{3\ \text{опт}}$, то обнулить список, и для данного запроса не показывать рекламный блок над результатами поиска. Иначе показать все рекламные объявления из усеченного списка, т.е. положить $t_{aq\ \text{опт}} = 1$. Остальные значения t_{aq} обнулить.

В этой возможности определить $t_{aq\ \text{опт}}$, не обращая внимания на другие запросы, и есть смысл декомпозиции.

2.3 Формальное описание алгоритма подбора параметров критерия показа.

Теперь запишем общий алгоритм подбора параметров критерия показа рекламных объявлений над результатами поиска.

Обозначения в алгоритме:

$\boldsymbol{Q}$ – общее число запросов;

$\boldsymbol{A}$ – общее количество рекламных объявлений-кандидатов на показ над результатами поиска;

$\boldsymbol{H_{max}}$ – максимальное допустимое покрытие,

$\boldsymbol{M_{min}}$ – минимальные денежные средства, которые должны быть выручены от показа рекламных объявлений над результатами поиска по нашей выборке запросов.

Для каждого запроса и объявления-кандидата дано:

$\boldsymbol{CTR_{aq}}$ – прогноз вероятности клика по рекламному объявлению

$\boldsymbol{Bid_a}$ – ставка рекламодателя по a –тому рекламному объявлению.

Цикл по λ_1 (например от 0 вверх)

Вырученные денежные средства с ростом λ_1 не убывают, то есть можно сказать что λ_1 – параметр, регулирующий поступление денежных средств от рекламодателей.

Цикл по E_0

$\boldsymbol{E_0}$ – общее количество показов рекламы над результатами поиска по всей выборке запросов.

Цикл по λ_2

Величина λ_2 регулирует количество показов в рекламном блоке по одному запросу.

Цикл по запросам j

По всем объявлениям-кандидатам a для запроса q:

Вычислить:

$$F_{aq} = CTR_{aq}/E_0 \ + \lambda_1 \cdot Bid_a \cdot CTR_{aq} - \lambda_2$$

Если $\boldsymbol{F_{aq} > 0}$, то включить пару (a, q) в -ый список:

Если в -ом списке объектов меньше чем k, то включить объявление a в список, иначе попытаться вытеснить худшее по F_{aq} и вставить a-ое, а если текущее объявление хуже всех по критерию F_{aq}, то не включать его.

Вычислить вклад в суммарный критерий объявлений, вошедших в список для показа:

$$R_q = \sum_{\text{(по объявлениям,вошедших в список)}} F_{aq}$$

Конец цикла по запросам q.

Упорядочить списки для запросов в порядке ***убывания*** $\boldsymbol{R_q}$.

Оставить только первые $\boldsymbol{H_{max}}$ из них (выполнение ограничения по покрытию), остальные обнулить (это в том

случае, если списков с рекламой хватает, в противном случае оставить все списки).

Запомнить $\boldsymbol{\lambda_3} = \min\limits_{q} R(q)$

Положить:

$$t_{aq} = \begin{cases} 1, & \text{если пара } (a, q) \text{ включена в список для показа,} \\ 0, & \text{в противном случае} \end{cases}$$

Вычислить $\boldsymbol{Events} = \sum_{(a,q)} t_{aq}$, при этом величина $Events$ будет зависеть от параметров (λ_1, λ_2, E_0).

Менять $\boldsymbol{\lambda_2}$, пока не будет достигнуто $Events = E_0 \pm \varepsilon$.

Положить $\lambda_{2\,\text{опт}} = \lambda_2$.

Запомнить λ_3, $\lambda_{2\,\text{опт}}$, списки t_{aq} (то есть объявления, отобранные для показа)

Конец цикла по $\boldsymbol{\lambda_2}$

Вычислить

$$CRIT_1(\lambda_1, \lambda_{2\,\text{опт}}, E_0) = \sum_{(a,q)} \boldsymbol{t}_{aq} \cdot \left(\frac{CTR_{aq}}{E_0} - \lambda_1 \cdot \left(M_{min} - M(\boldsymbol{T}) \right) \right)$$

Менять $\boldsymbol{E_0}$, чтобы достигнуть $\max\limits_{E_0} CRIT_1$

Положить $E_{0\,\text{опт}}$ то значение E_0 , при котором достигнут $\max\limits_{E_0} CRIT_1$.

Запомнить λ_3 , $\lambda_{2\,\text{опт}}$, $E_{0\,\text{опт}}$, $M(\boldsymbol{T})$

Конец цикла по $\boldsymbol{E_0}$

Менять $\boldsymbol{\lambda_1}$ пока $M(\boldsymbol{T}) < M_{min}$

Конец цикла по $\boldsymbol{\lambda_1}$

В конце работы алгоритма мы получаем четыре величины $\lambda_{1\,\text{опт}}$, $E_{0\,\text{опт}}$, $\lambda_{2\,\text{опт}}$ и $\lambda_{3\,\text{опт}}$, которые будут использоваться для работы с новыми запросами.

2.4 Работа с новыми запросами.

Рассмотрим то, каким образом мы будем использовать полученные с помощью нашего алгоритма параметры $\lambda_{1\,\text{опт}}$, $E_{0\,\text{опт}}$, $\lambda_{2\,\text{опт}}$ для обработки нового запроса, поступившего от пользователя в систему показов рекламных объявлений.

Ранее было отмечено, что если данные параметры критерия показа уже получены, то для каждого запроса из выборки запросов можно определить, какие рекламные объявления должны быть показаны в рекламном блоке над результатами поиска, не обращая внимания на другие запросы. Мы рассчитываем на то, что используемая выборка запросов достаточно велика для того чтобы результат подбора параметров критерия показа мог быть применим в тех же условиях, в которых собиралась эта выборка. Тогда, в силу закона больших чисел, если мы будем применять те же правила к новым запросам, то интегральные критерии – среднее значение кликабельности, суммарные денежные средства и покрытие будут достаточно близки к тому, что было получено по выборке запросов. В то же время параметры были подобраны так, чтобы наш главный критерий – среднее значение CTR по рекламным показам над результатами поиска достиг максимума при заданных ограничениях на суммарный доход и покрытие. Поэтому и на новых данных с достаточной точностью в среднем критерий будет достигать максимума, а ограничения выполняться.

<u>Итак, получив новый запрос нужно:</u>

1) Отобрать объявления-кандидаты для возможных показов (по фразам запроса).
2) Для каждого из отобранных объявлений известно значение ставки Bid_a и прогноз вероятности клика по объявлению CTR_{aq}, где a – индекс баннера, а q – индекс запроса.

3) Вычислить значение $F_{aq} = CTR_{aq}/E_{0\ \text{опт}} + \lambda_{1\ \text{опт}} \cdot Bid_a \cdot CTR_{aq} - \lambda_{2\ \text{опт}}$ для всех полученных объявлений.
4) Составить список кандидатов на показ над результатами поиска. В этот список для показа поместить только те объявления, для которых $CTR_{aq}/E_{0\ \text{опт}} + \lambda_{1\ \text{опт}} \cdot Bid_a \cdot CTR_{aq} > \lambda_{2\ \text{опт}}$. Этот список может быть и пустым (в этом случае над результатами поиска не будет показано ни одного рекламного объявления).
5) Если число кандидатов больше k, то сократить список, оставив в нем только k объявлений с наибольшим значением F_{aq}.
6) Подсчитать значение $R_q = \sum_{(a)}(CTR_{aq}/E_{0\ \text{опт}} + \lambda_{1\ \text{опт}} \cdot Bid_a \cdot CTR_{aq} - \lambda_{2\ \text{опт}})$, где сумма берется по всем объявлениям из усеченного списка.
7) Если эта величина меньше $\lambda_{3\ \text{опт}}$, то обнулить список, и для данного запроса не показывать никаких объявлений над результатами поиска. Иначе показать все объявления из усеченного списка.

После проведения операций 1)-7) , будет совершенно ясно показывать ли по новому поступившему запросу объявления, если да, то какие конкретно из всех объявлений-кандидатов на показ.

Таким образом, представлена математическая постановка задачи оптимизации показов объявлений над результатами поиска, а также решение этой оптимизационной задачи при помощи теории Куна-Таккера. В итоге представлен алгоритм отбора рекламных объявлений для показа, а также обработка нового пользовательского запроса поступившего в поисковую систему.

2.5 Модификация алгоритма

Предлагается модификация базового алгоритма, позволяющая существенно ускорить его работу. Общая схема алгоритма представлена на Рис. 3.

Внешний цикл реализует перебор по переменной λ_2, в ходе которого ищется значение $\lambda_{2\text{опт}}$, доставляющее максимум нашему основному критерию:

$$CRIT_0(\boldsymbol{T}) \;\; = \sum_{(a,q)} CTR_{aq} \cdot t_{aq}/Events$$

вложенный цикл осуществляет перебор по переменной λ_1, где ищется (при фиксированном значении λ_2) значение λ_1, обеспечивающее равенство $M(\boldsymbol{T}) = M_{min}$.

Замена неравенства $M(\boldsymbol{T}) \geq M_{min}$ на равенство объясняется следующим: если бы было справедливо $M(\boldsymbol{T}) > M_{min}$, то по теории Куна-Таккера должно быть справедливо $\lambda_1 = 0$, но тогда ограничение на суммарные денежные средства вообще не учитывается, и максимум среднего значения CTR будет достигнут, если показать только объявления с максимальным значением CTR. Но таковых будет один или совсем немного, а тогда (в практически интересных случаях) денежных средств заведомо не хватит.

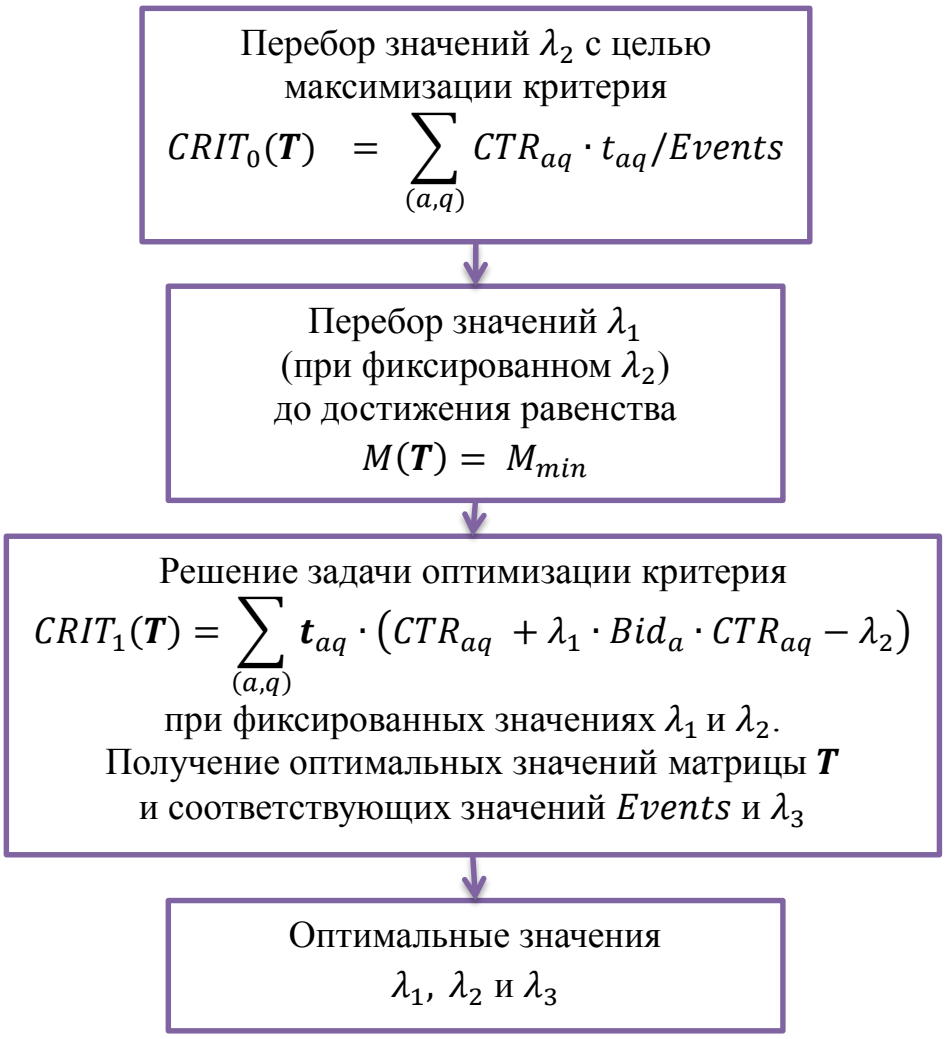

Рис. 3. Общая схема алгоритма подбора порогов.

Можно показать, что оптимальная величина $M(\boldsymbol{T})$ (при фиксированном значении λ_2) не убывает с ростом λ_1, и поэтому для нахождения корня $M(\boldsymbol{T}) = M_{min}$ можно использовать бинарный поиск.

Центральная часть алгоритма осуществляет выбор оптимальных значений переменных t_{aq}, доставляющих максимум критерию:

$$CRIT_1(\boldsymbol{T}) = \sum_{(a,q)} t_{aq} \cdot (CTR_{aq} + \lambda_1 \cdot Bid_a \cdot CTR_{aq} - \lambda_2)$$

при фиксированных значениях λ_1 и λ_2. Оптимизация идет тем же путем, который описан для начального алгоритма.

2.5.1 Формальное описание алгоритма

Теперь запишем общий алгоритм подбора параметров критерия показа объявлений над результатами поиска. Обозначения приняты такими же, как и в формальном описании начального алгоритма (п. 2.3).

Алгоритм можно записать в виде:

Цикл по λ_2

Величина λ_2 доставляет максимум основному критерию $CRIT_0(\boldsymbol{T})$.

Цикл по λ_1

Вырученные денежные средства с ростом λ_1 не убывают, то есть можно сказать что λ_1- параметр, регулирующий поступление денежных средств от рекламодателей.

Перебор по λ_1 идёт до достижения $M(\boldsymbol{T}) = M_{min}$

Цикл по запросам q

По всем баннерам-кандидатам a для запроса q:

Составить «список для показа», вычисляя для каждого рекламного объявления:

$$F_{aq}(\lambda_1, \lambda_2, E_0) = CTR_{aq} + \lambda_1 \cdot Bid_a \cdot CTR_{aq} - \lambda_2$$

Не включая в список те a, для которых F_{aq} не положительно, и оставив k объявлений с наибольшим значением F_{aq}.

Вычислить вклад в суммарный критерий объявлений, вошедших в список для показа:

$$R_q = \sum_{\text{(по объявлениям вошедшим в список)}} F_{aq}$$

Конец цикла по запросам q

Упорядочить списки для запросов в порядке убывания R_q.

Оставить только первые H_{max} из них (выполнение ограничения по покрытию), остальные обнулить (это в том случае, если списков с рекламой хватает, в противном случае оставить все списки).

Запомнить $\boldsymbol{\lambda_3} = \min_q R_q$

Положить:

$$t_{aq} = \begin{cases} 1, & \text{если пара } (a,q) \text{ оставлена для показа,} \\ 0, & \text{в противном случае} \end{cases}$$

Менять λ_1, пока не будет достигнуто $MO(\boldsymbol{T}) = M_{min}$ следующим образом:

Если $M(\boldsymbol{T}) < M_{min}$, то уменьшить λ_1

Если $M(\boldsymbol{T}) > M_{min}$, то увеличить λ_1

Положить $\boldsymbol{\lambda_{1\,\text{опт}}} = \lambda_1$.

Запомнить λ_3, $\lambda_{1\,\text{опт}}$, списки t_{aq} (то есть объявления, отобранные для показа)

Конец цикла по $\boldsymbol{\lambda_1}$.

Менять λ_2 чтобы достигнуть ***максимума***

$$CRIT_0(\boldsymbol{T}) = \sum_{(a,q)} CTR_{aq} \cdot t_{aq}/Events$$

Положить $\boldsymbol{\lambda_{2\text{опт}}}$ то значение λ_2, при котором достигнут $\max CRIT_0(\boldsymbol{T})$.

Конец цикла по $\boldsymbol{\lambda_2}$

В конце работы алгоритма мы получаем значения трёх параметров $\lambda_{1\,\text{опт}}$, $\lambda_{2\,\text{опт}}$ и $\lambda_{3\,\text{опт}}$, которые будут использоваться для работы с новыми запросами.

2.5.2 Доказательство эквивалентности двух алгоритмов

Допустим, что максимум функционала:

$$CRIT_0(\boldsymbol{T}) = \sum_{(a,q)} CTR_{aq} \cdot t_{aq}/Events$$

достигается в $\boldsymbol{T}_{\text{опт}}$ и величина $Events_{\text{опт}}$ соответствует этому значению $\boldsymbol{T}_{\text{опт}}$, где $t_{aq} = 0$ означает, что рекламное объявление на запрос не показывалось, а при $t_{aq} = 1$ – показывалось, и $Events = \sum_{(a,q)} t_{aq}$, при ограничениях:

$$M(\boldsymbol{T}) \geq M_{min}$$

$$\forall q \sum_{a} t_{aq} \leq k$$

$$H(\boldsymbol{T}) = \sum\nolimits_{(q)} \mathbb{I}\left\{\sum\nolimits_{(a)} t_{aq} > 0\right\} \leq H_{max}$$

$$0 \leq t_{aq} \leq 1$$

В дальнейшем предположим, что в точке максимума ограничение на суммарные денежные средства достигается, т.е.

$$\boldsymbol{M}(\boldsymbol{T}_{\text{опт}}) = \boldsymbol{M_{min}} \qquad (7)$$

По «старому» алгоритму $T_{\text{опт}}$ достигается при следующих условиях: Критерий

$$CRIT_2(\boldsymbol{T}) = \sum_{(a,q)} t_{aq} \cdot (CTR_{aq}/E_0 + \lambda_1 \cdot Bid_a \cdot CTR_{aq} - \lambda_2) + const$$

достигает максимума по $\boldsymbol{T}$ при ограничениях:

$$\forall q \sum_{a} t_{aq} \leq k$$

$$H(\boldsymbol{T}) = \sum\nolimits_{(q)} \mathbb{I}\left\{\sum\nolimits_{(a)} t_{aq} > 0\right\} \leq H_{max}$$

$$0 \leq t_{aq} \leq 1$$

а коэффициенты λ_1, λ_2 и E_0 подбираются так, чтобы

$$Events = \sum_{(a,q)} t_{aq} = E_0$$

$$E_{0_{\text{опт}}} = \arg\max CRIT_1(\boldsymbol{T}, \lambda_1, E_0) =$$

$$= \arg\max\left[\frac{\sum CTR_{aq} \cdot t_{aq}}{E_0} - \lambda_1\left(M_{min} - M(\boldsymbol{T})\right)\right]$$

$$M(\boldsymbol{T}_{\text{опт}}) = M_{min},$$

Следовательно

$$E_0 = Events_{\text{опт}}.$$

Обозначим также подобранные значения λ_1 и λ_2 как $\lambda_{1\text{опт}}$ и $\lambda_{2\text{опт}}$.

Умножим $CRIT_2(\boldsymbol{T})$ на постоянную величину $Events_{\text{опт}}$ и обозначим эту функцию $CRIT_2^*(\boldsymbol{T})$:

$$CRIT_2^*(\boldsymbol{T}) = \sum_{(a,q)} t_{aq} \cdot (CTR_{aq} \ + (\lambda_1 \cdot Events_{\text{опт}}) \cdot Bid_a \cdot CTR_{aq} - (\lambda_2 \cdot Events_{\text{опт}})) \ + const$$

Замечание 1.

Обозначим $\lambda_1^* = \lambda_1 \cdot Events_{\text{опт}}$, $\lambda_2^* = \lambda_2 \cdot Events_{\text{опт}}$, и заметим, что $T_{\text{опт}}$, доставляющее максимум критерию, не меняется при умножении на константу.

Перейдем теперь к «новому алгоритму». По этому алгоритму мы сначала максимизируем критерий $\sum_{(a,q)} t_{aq} \cdot (CTR_{aq} \ + \lambda_1 \cdot Bid_a \cdot CTR_{aq} - \lambda_2)$ при фиксированных значениях λ_1 и λ_2. При этом в качестве оптимальных значений получим $\boldsymbol{T}(\lambda_1, \lambda_2)$ и соответствующие значения $Events(\lambda_1, \lambda_2)$, $CRIT_0\left(\boldsymbol{T}(\lambda_1, \lambda_2)\right)$ и $M\left(\boldsymbol{T}(\lambda_1, \lambda_2)\right)$**.** Затем для каждого значения λ_1 величина $\lambda_{2_{\text{опт}}}(\lambda_1)$ подбирается так, чтобы

$$M(\lambda_1, \lambda_2) = M_{min}$$

Далее значение $\lambda_2 = \lambda_{2_{\text{опт}}}$ подбирается так, чтобы величина $CRIT_0\left(\boldsymbol{T}(\lambda_1, \lambda_2)\right)$ достигла максимума при $\lambda_2 = \lambda_{2_{\text{опт}}}(\lambda_1)$.

Теперь видим, что значения λ_1^* и λ_2^* могут претендовать на результат «нового» алгоритма. Действительно, в силу **Замечания 1** значение $\boldsymbol{T}(\lambda_1^*, \lambda_2^*) = \boldsymbol{T}_{\text{опт}}$, и, значит, доставляет максимум критерию

$\sum_{(a,q)} t_{\text{опт}} \cdot (CTR_{aq} + \lambda_1 \cdot Bid_q \cdot CTR_{aq} - \lambda_2)$ при фиксированных значениях λ_1^* и λ_2^*. В то же время в силу (1):

$$M(\lambda_1^*, \lambda_2^*) = M_{min}$$

Таким образом, при переборе λ_1 в качестве нужного значения при $\lambda_1 = \lambda_1^*$ будет по новому алгоритму выбрано $\lambda_2 = \lambda_2^*$. Наконец, по новому алгоритму значение $\lambda_{1_{\text{опт}}}$ выбирается так, чтобы был достигнут максимум величины $CRIT_0\big(\boldsymbol{T}(\lambda_1, \lambda_2)\big)$. Но при $\lambda_1 = \lambda_1^*$ и $\lambda_2 = \lambda_2^*$ эта величина по построению достигает максимума. Значит, либо $\lambda_{1_{\text{опт}}} = \lambda_1^*$, либо при равенстве максимумов

$$CRIT_0\big(\boldsymbol{T}(\lambda_1^*, \lambda_2^*)\big) = CRIT_0\left(\boldsymbol{T}\big(\lambda_{1_{\text{опт}}}, \lambda_{2_{\text{опт}}}\big)\right).$$

И, значит, новый алгоритм дает результат равносильный старому.

Таким образом, модифицированный алгоритм решает ту же оптимизационную задачу, однако количество затрачиваемого объёма производимых операций уменьшается на порядок (так как был упразднён один из переборов по внешнему параметру).

2.6 Новая модель показа рекламных объявлений.

Теперь, когда получен новый вид критерия показа рекламного объявления в рекламном блоке над результатами поиска, можно описать новую модель показа.

Входные данные модели остаются такими же: набор запросов пользователей q ($1 \le q \le Q$). Для каждого из запросов составляется список рекламных объявлений a ($1 \le a \le A$) для которых известно: предсказание вероятности клика CTR_{aq} и ставка Bid_a.

Процесс моделирования. Имеется набор запросов и набор рекламных объявлений и необходимо решить какие конкретно из объявлений следует показать. Был выявлен вид критерия показа $CTR_{aq} + \lambda_1 \cdot Bid_a \cdot CTR_{aq} - \lambda_2 > 0$ с помощью которого модель

определяет показывать ли объявление a по запросу q. Для каждого из запросов и соответствующим им объявлениям необходимо выполнить:

1) Вычислить значение $F_{aq} = CTR_{aq} + \lambda_{1\,\text{опт}} \cdot Bid_a \cdot CTR_{aq} - \lambda_{2\,\text{опт}}$ для всех объявлений.
2) Составить список кандидатов на показ над результатами поиска. В этот список поместить только те объявления, для которых $F_{aq} > 0$. Если список пуст, то над результатами поиска не будет показано ни одного рекламного объявления.
3) Если число кандидатов больше k, то сократить список, оставив в нем только k объявлений с наибольшим значением F_{aq}.
4) Подсчитать значение $R_q = \sum_{(a)} F_{aq}$, где сумма берется по всем объявлениям из усеченного списка.
5) Если эта величина R_q меньше $\lambda_{3\,\text{опт}}$, то обнулить список, и для данного запроса не показывать никаких объявлений над результатами поиска и положить соответствующие $t_{aq} = 0$. Иначе показать все объявления из усеченного списка и положить $t_{aq} = 1$.

Таким образом заполняется матрица $\boldsymbol{T}$ показов по всему набору запросов Q и множеству рекламных объявлений A.

Выходные характеристики системы. После того как матрица T определена и известно какие объявления покажутся по каждому из запросов считаются суммарные характеристики системы: суммарный доход $M(\boldsymbol{T})$, средняя кликабельность по всему набору запросов $CTR_{avg}(\boldsymbol{T})$ и количество запросов с показанным по ним рекламным блоком $H(\boldsymbol{T})$. Теперь, они соответствуют следующим правилам:

$$CTR_{avg}(\boldsymbol{T}) \to \max$$

$$M(\boldsymbol{T}) \geq M_{min}$$

$$H(\boldsymbol{T}) \leq H_{max}$$

$$\forall\, q\colon\ k_q \leq\ k$$

То есть мы отбираем объявления для показа таким образом чтобы максимизировать среднюю кликабельность всех показанных объявлений по набору запросов с ограничениями на суммарный доход и покрытие.

Достоинством данной модели (Рис. 4.) является то, что для управления выходными характеристиками системы показов рекламных объявлений используется три параметра критерия показа λ_1, λ_2 и λ_3. С помощью этих параметров мы можем регулировать основные показатели системы.

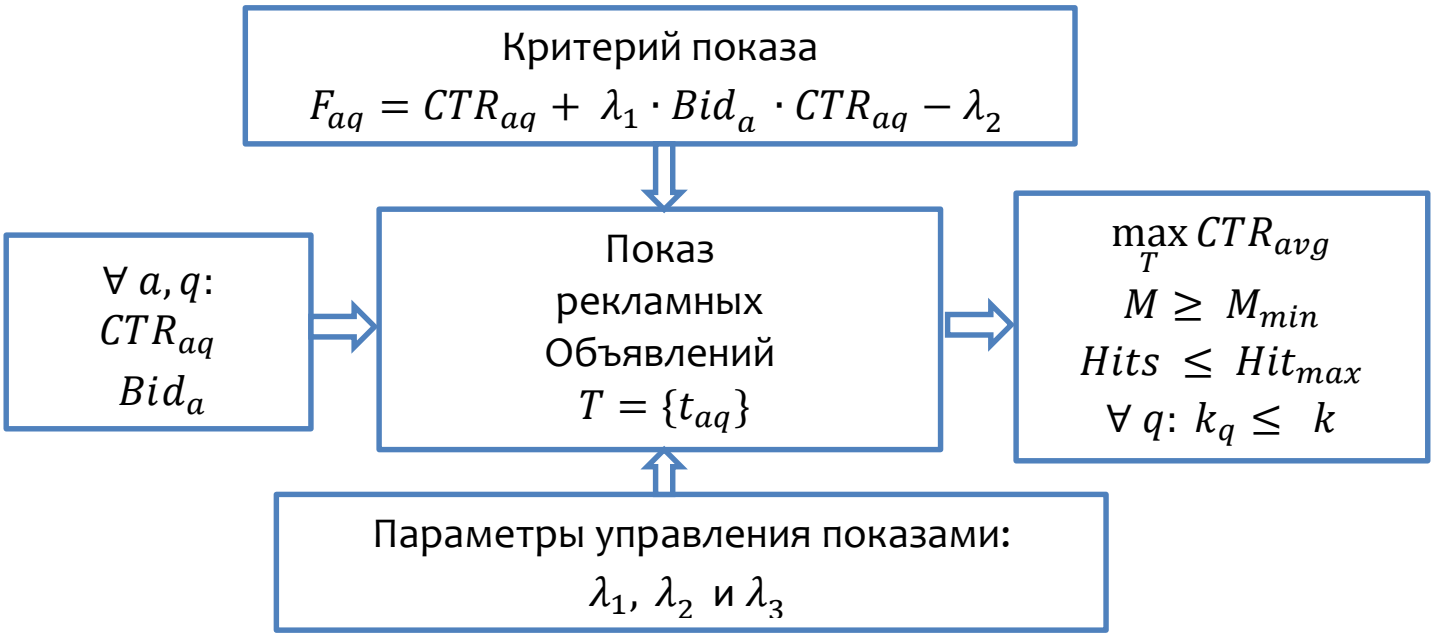

Рис. 4. Новая модель показов рекламы над результатами поиска.

2.7 Результаты экспериментального тестирования алгоритма

Тестирование на реальных данных будет проводиться только для улучшенного алгоритма по следующим причинам:

во-первых, модифицированный алгоритм эквивалентен изначально предложенному (п. 2.5.2);

во-вторых, уменьшено машинное время работы алгоритма, поэтому применять его к большим объёмам данных более целесообразно;

в-третьих, при варьировании меньшего количества параметров, гораздо проще проводить интерпретацию работы алгоритма;
и, наконец, ***в-четвёртых*** более «прозрачная» схема работы позволяет лучше понять по каким критериям происходит отбор объявлений для показа в рекламном блоке над результатами поиска.

2.7.1 Данные для тестирования

Экспериментальное тестирование описанного выше алгоритма будет проводиться следующим образом: последовательно рассмотрим каждый из этапов выбора значений параметров λ_1, λ_2 и λ_3 и убедимся, что алгоритм отбирает объявления нужным нам способом.

Данные для тестирования выбирались следующим образом:

1) Бралась неделя логов показов рекламы на поисковой странице компании «Яндекс» за 25 – 31 января 2013 г.
2) Из данных по поисковому логу выделялись запросы, которые пользователи задавали за соответствующую неделю.
3) Для тестирования алгоритма отбиралось 100 000 запросов (случайно, то есть выборка запросов получилась несмещённая и репрезентативная).
4) Из запроса выделяется ключевая фраза с помощью морфологического инструментария. Фраза, точно (с точностью до словоформы и предлогов/союзов), совпадающая с запросом называется **точным совпадением**. Также существует большое количество способов подбора дополнительных фраз [37]. Таким образом, каждому запросу ставится в соответствие набор фраз.
5) Каждой фразе запроса соответствует множество рекламных объявлений, которые торгуются за попадание в рекламный показ (и в том числе за позицию при показе) по этой фразе. Для запроса

составляется список рекламных объявлений как объединение их множеств по каждой из фраз, соответствующих запросу.

6) Для каждого объявления известно:

   CTR_{aq} – предсказание вероятности клика для пары объявление-фраза.

   Bid_a – ставка для данного баннера и соответствующей фразы.

2.7.2 Этапы проведения экспериментального тестирования.

Экспериментальное тестирование алгоритма будет состоять из следующих этапов:

Этап 1. Рассмотрим один запрос.

Рассмотрим ту часть алгоритма, которая на схеме (Рис. 2.) обозначается (Рис. 5.).

Решение задачи оптимизации критерия

$$CRIT_1(\boldsymbol{T}) = \sum_{(a,q)} \boldsymbol{t}_{aq} \cdot \left(CTR_{aq} + \lambda_1 \cdot Bid_a \cdot CTR_{aq} - \lambda_2\right)$$

при фиксированных значениях λ_1 и λ_2.
Получение оптимальных значений матрицы $\boldsymbol{T}$
и соответствующих значений $Events$ и λ_3

Рис. 5. Работа с отдельными запросами.

Сначала возьмём один запрос и, перебирая λ_1, зафиксируем как последовательно отбираются объявления на показ в рекламный блок над результатами поиска. Рассматриваются те характеристики, которые задействованы в задаче оптимизации (5).

Этап 2. Подбор λ_1 для фиксированного λ_2.

Перебор значений λ_1
(при фиксированном λ_2)
до достижения равенства
$M(\boldsymbol{T}) = M_{min}$

Рис. 6. Подбор значения параметра $\boldsymbol{\lambda_1}$ при фиксированном $\boldsymbol{\lambda_2}$ для всего набора запросов.

Возьмём полный набор запросов с объявлениями – кандидатами. Для заданного λ_2 варьируем λ_1 до достижения равенства по прогнозируемым денежным средствам (Рис. 6.)

Для пары параметров $\boldsymbol{\lambda_1}$ и $\boldsymbol{\lambda_2}$ уже становится возможным подсчёт суммарных характеристик для всего набора запросов (то есть считаются уже характеристики всей системы). Сам параметр λ_1 отвечает за вклад предсказанного дохода от показа рекламного объявления, тем самым мы регулируем и суммарный ожидаемый доход, который входит в (5) как ограничение. Нас интересует зависимость от λ_1 таких величин как: $CTR, Bid \cdot CTR, Events/\ Hits, \lambda_3$.

Этап 3. Максимизация основного критерия.

Внешний цикл алгоритма отвечает за оптимизацию основного критерия оптимизационной задачи – среднего CTR по всему набору запросов (Рис. 7.).

Перебор значений λ_2 с целью
максимизации критерия
$$CRIT_0(\boldsymbol{T}) = \sum_{(a,q)} CTR_{aq} \cdot t_{aq}/Events$$

Рис. 7. Максимизация основного критерия.

Будем варьировать λ_2, получать для него соотвествующее оптимальное λ_1 (которое выводит систему на ограничение по ожидаемому доходу), и получать для пары (λ_1, λ_2) набор различных величин $CTR, Bid \cdot CTR, CRIT_0$.

2.7.3 Отбор объявлений на показ для одного запроса.

Было взято фиксированное значение $\lambda_2 = 7.46$, λ_1 варьировалось с шагом 0.1. Зафиксировав λ_2, и варьируя λ_1 (последовательно увеличивая) мы начинаем «пускать» объявления для показа в рекламном блоке над результатами поиска. Рассмотрим на примере как это происходит. Допустим, в качестве кандидатов на показ у нас отобрались объявления со следующими характеристиками (Табл.1.)

ID	CTR	Bid	CPM
1	0.0452	100	4.52
2	0.0271	122	3.30
3	0.0952	10	0.95
4	0.0284	66	1.87
5	0.0534	11	0.59
6	0.0627	3	0.19
7	0.0254	18	0.46

Табл. 1. Набор характеристик для объявлений-кандидатов для показа по конкретному запросу.

Для первой пары $(\lambda_1, \lambda_2) = (1,7.46)$ получаются следующие значения $CRIT_1$ (Табл.2.).

ID	$\boldsymbol{CRIT_1}$
1	-1.83
2	-3.49
3	-4.16
4	-4.89
5	-5.56
6	-5.73
7	-6.38

Табл. 2. Значение критерия $\boldsymbol{CRIT_1}$ для объявлений-кандидатов.

Из таблицы видно, что для всех объявлений значение $CRIT_1$ отрицательно, и ни одно объявление не будет претендентом на показ.

Теперь, последовательно увеличивая λ_1, можно найти такое λ_1, чтобы показать только одно объявление, это $\lambda_1 = 1.5$. Будем искать следующее значение λ_1 такое, что в список кандидатов на показ в

рекламном блоке добавляется ещё одно дополнительное объявление. Получены соответствующие значения λ_1 и $CRIT_1$ для объявлений-кандидатов (Табл. 3.**)**.

ID	$CRIT_1$ $(\lambda_1 = 1.5)$	$CRIT_1$ $(\lambda_1 = 2.1)$	$CRIT_1$ $(\lambda_1 = 3.7)$	$CRIT_1$ $(\lambda_1 = 5.4)$
1	0.43	3.14	10.36	18.04
2	-1.84	0.14	5.42	11.03
3	-3.95	-2.83	0.16	3.35
4	-3.69	-3.12	-1.59	0.03
5	-5.26	-4.91	-3.97	-2.97
6	-5.63	-5.52	-5.14	-4.36
7	-6.15	-5.87	-5.22	-4.90

Табл. 3. Значение критерия $\boldsymbol{CRIT_1}$ для разных значений $\boldsymbol{\lambda_1}$.

Красным цветом отмечены положительные значения критерия, то есть когда объявление становится кандидатом на показ. По таблице (Табл. 3.) видно, что в последней колонке уже становится четыре кандидата на показ, однако покажется всего три (из-за ограничения на количество показываемых объявлений в рекламном блоке сверху результатов поиска на запрос пользователя). Покажем более наглядно динамику выбора алгоритмом объявлений (Рис. 8. – Рис. 12).

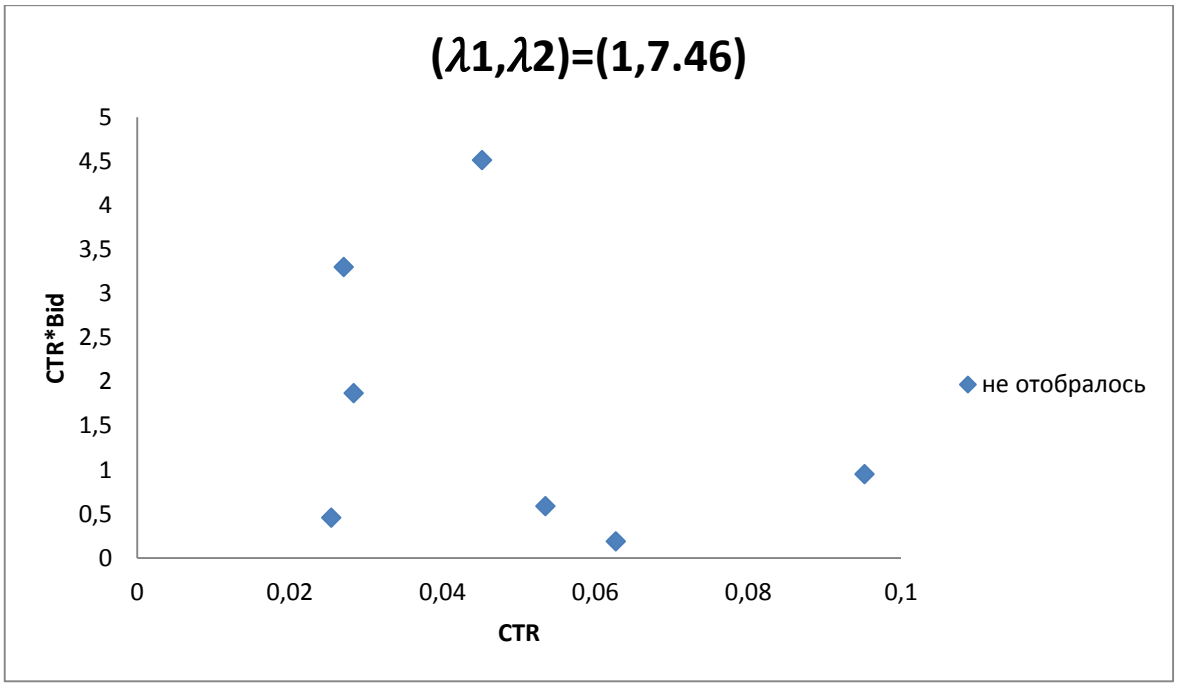

Рис. 8. Ни одного объявление для показа не отобралось.

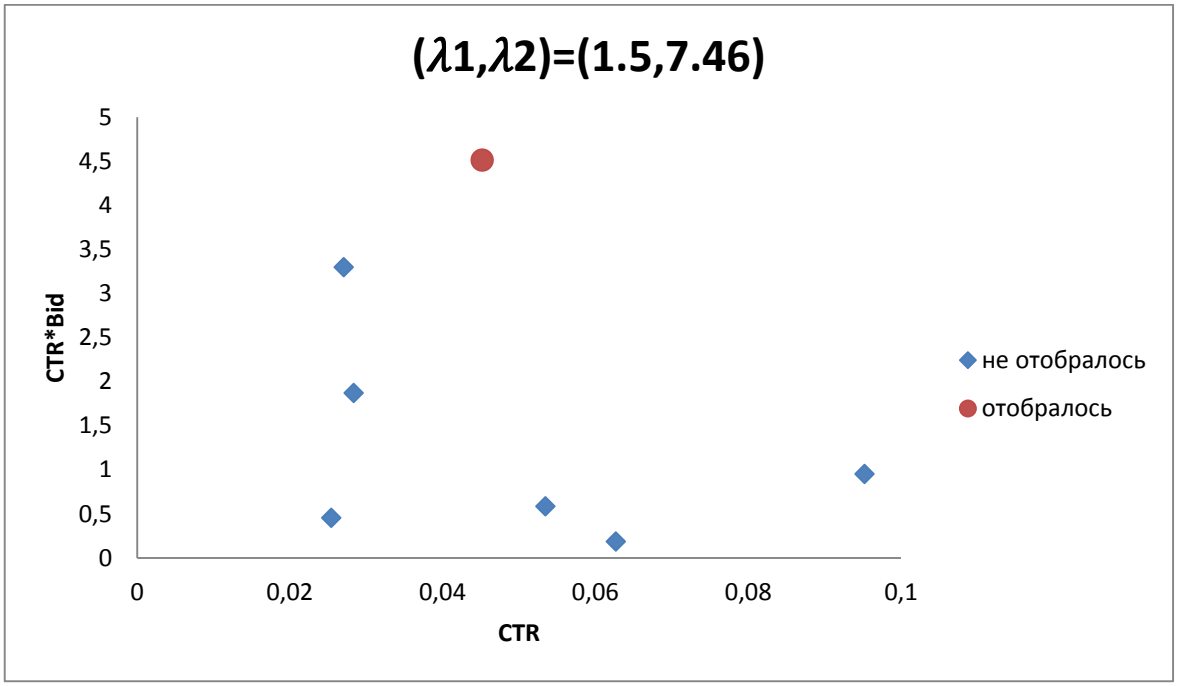

Рис. 9. Отобралось одно объявление на рекламный показ.

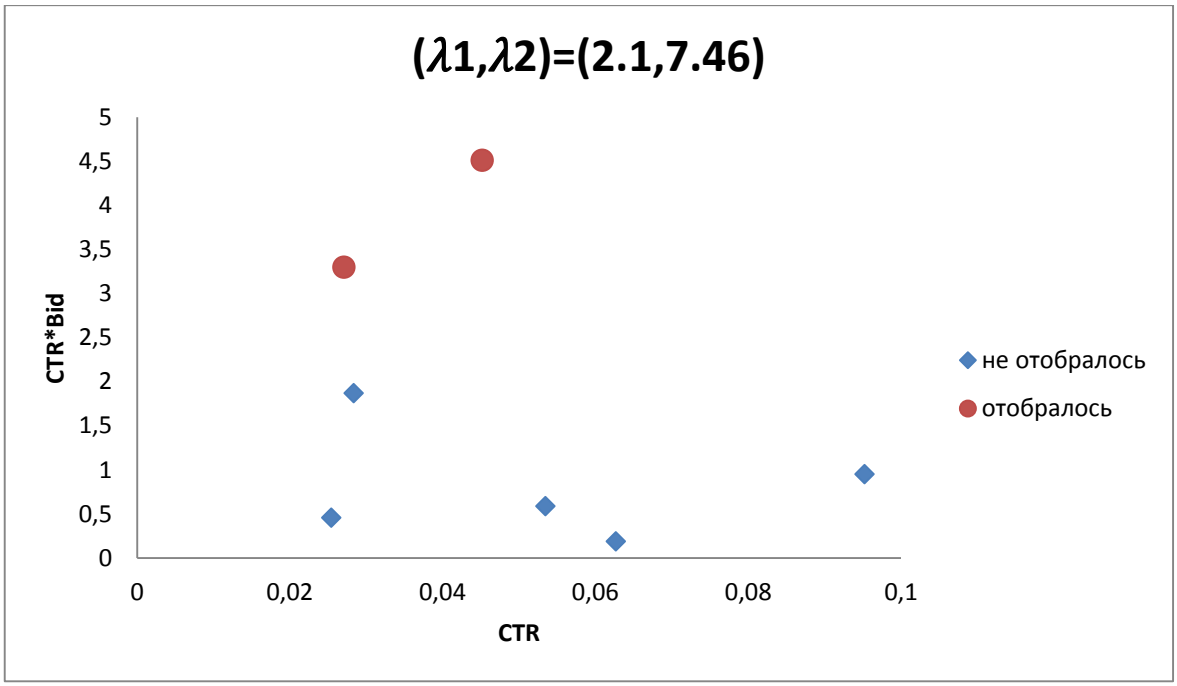

Рис. 10. Отобралось два объявления для показа.

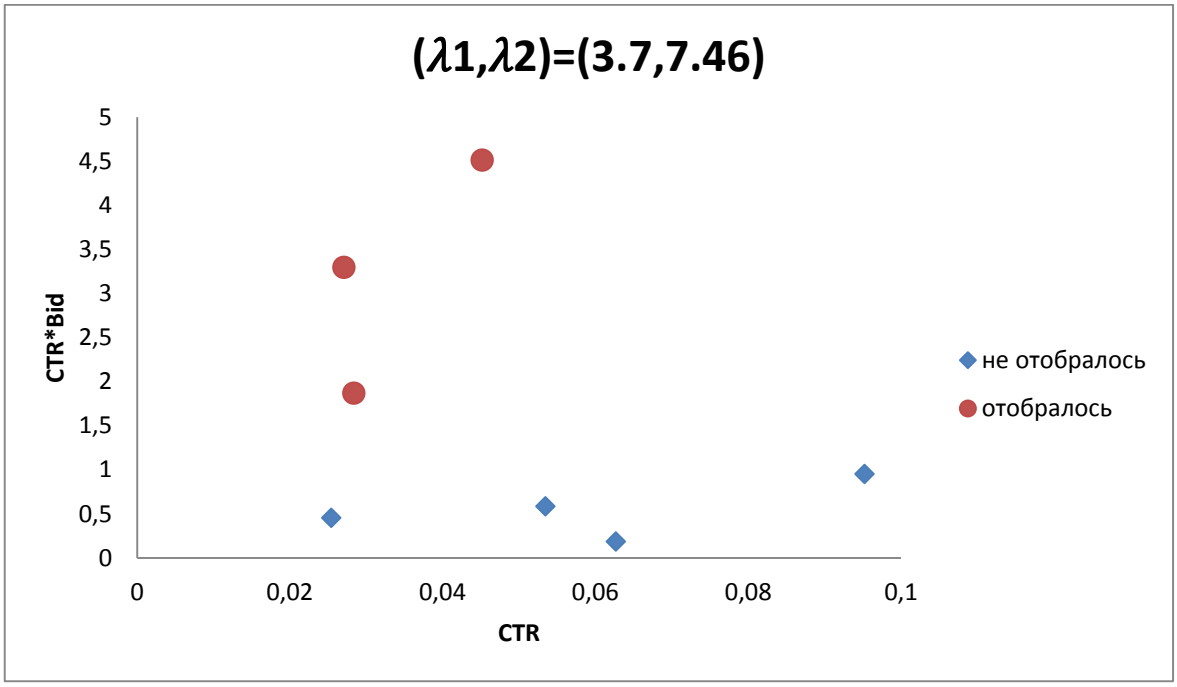

Рис. 11. Отобралось три объявления (максимально возможное значение)

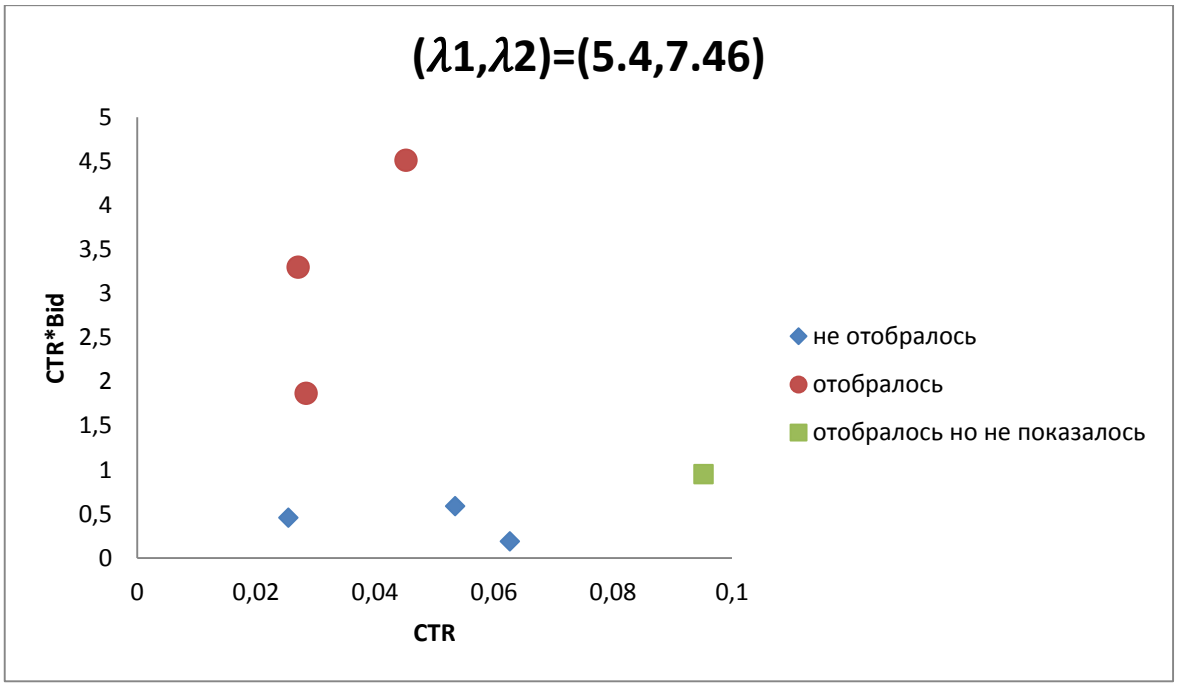

Рис. 12. По критерию отобралось четыре объявления, но будет показано всего три.

Таким образом, получено представление о том, как алгоритм отбирает из всего множества объявлений-кандидатов претендентов на показ в рекламном блоке для одного запроса.

2.7.4 Подбор параметра λ_1 при фиксированном значении λ_2 на всём пуле запросов.

Теперь будем варьировать λ_1 и λ_2 для всего набора запросов (одновременно) и посчитаем следующие величины: $CTR, Bid \cdot CTR$ vs λ_1. Рассмотрим зависимости основных показателей системы от параметра λ_1. Этот параметр входит в критерий показа следующим образом:

$$CTR_{aq} + \lambda_1 \cdot Bid_a \cdot CTR_{aq} - \lambda_2$$

То есть он присутствует перед $Bid \cdot CTR$– вероятностью списания ставки с рекламодателя. Рассмотрим поведение CTR и $Bid \cdot CTR$ от изменения параметра λ_1:

1) ***CTR vs λ_1.***

Получен график зависимости среднего CTR от λ_1 (Рис. 13.).

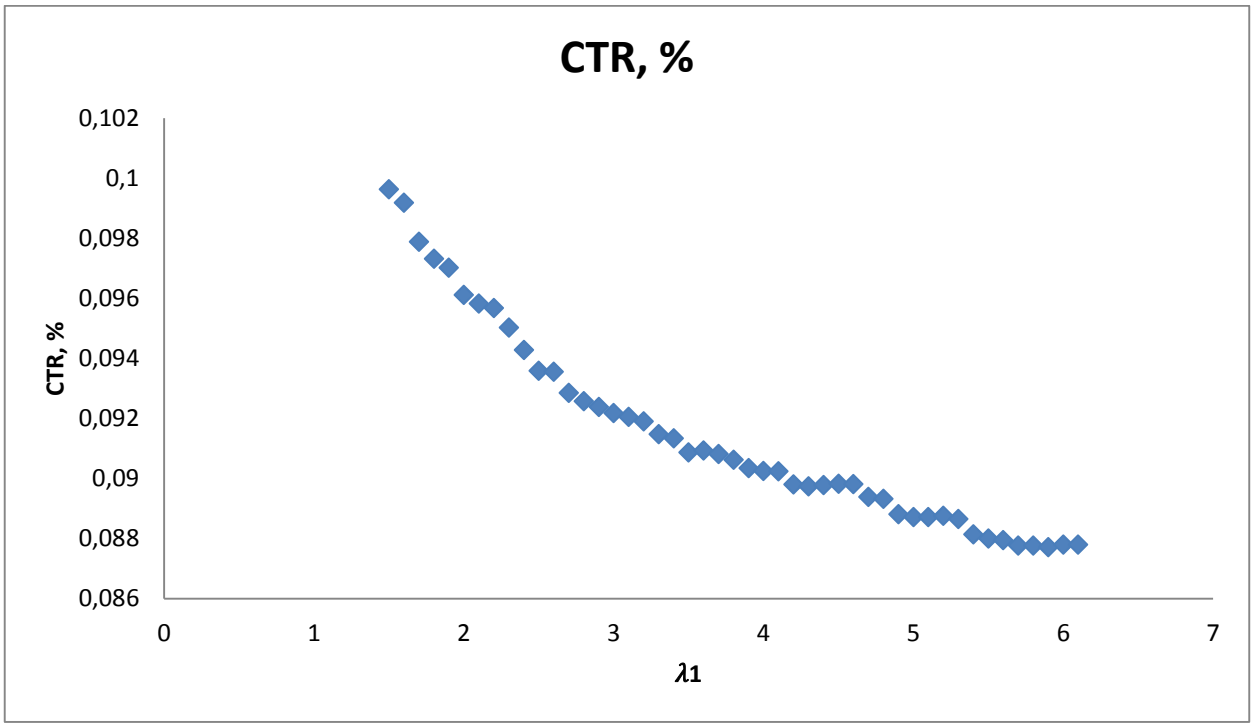

Рис. 13. Зависимость изменения среднего $\boldsymbol{CTR}$ от параметра $\boldsymbol{\lambda_1}$.

Из графика (Рис. 13.) видно, что, действительно, при увеличении λ_1 такая характеристика объявления как его CTR начинает учитываться при отборе объявления для показа всё меньше и меньше и средний CTR по всему набору запросов так же падает.

2) $\boldsymbol{Bid \cdot CTR\ vs\ \lambda_1}$.

Рассмотрим поведение $Bid \cdot CTR$ в зависимости от изменения параметра λ_1 (Рис. 14).

Опишем процедуру выбора $\lambda_{1\text{опт}}$ при фиксированом значении λ_2. Последовательно перебирая значения параметра λ_1, алгоритм получает суммарное значение $\sum Bid_{aq} \cdot CTR_a$, которое в нашем случае мы принимаем как аппроксимацию суммарного дохода $M(\boldsymbol{T})$. Ограничение оптимизационной задачи имеет вид: $M(\boldsymbol{T}) \geq M_{min}$. Как только M_{min} достигается, соотвествующее значение λ_1 для фиксированного λ_2 принимается за оптимальное и запоминается как $\lambda_{1\text{опт}}(\lambda_2)$.

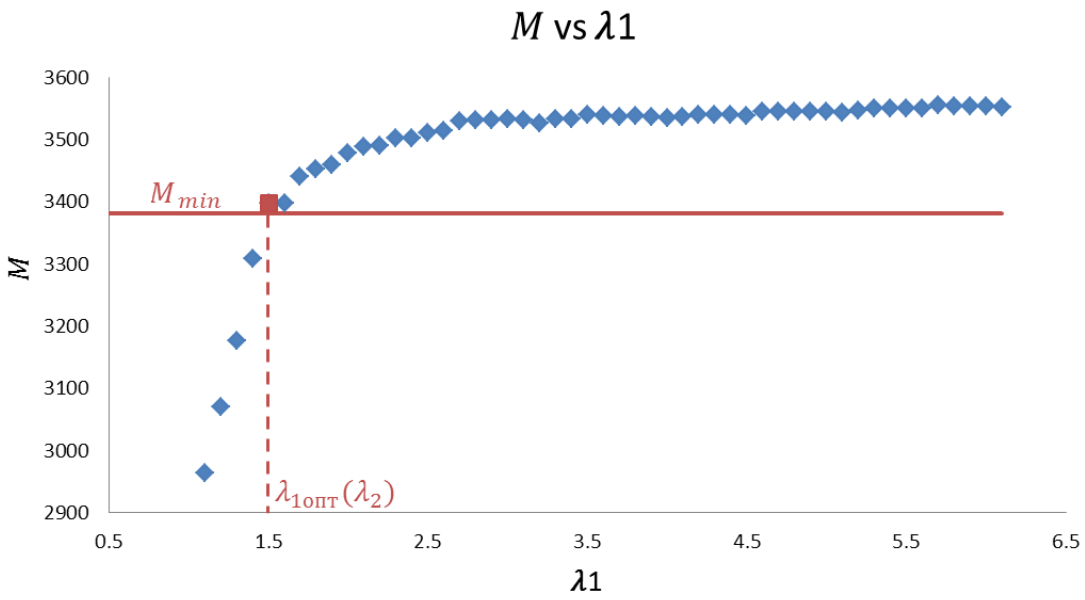

Рис. 14. Динамика изменения $\boldsymbol{M(T)}$ в зависимости от $\boldsymbol{\lambda_1}$.

На графике (Рис. 14) показана вся кривая варьирования λ_1 и $M(\boldsymbol{T})$, алгоритм перейдёт к следующей итерации по λ_2, как только будет достигнуто для пары (λ_1, λ_2) ограничение $M(\boldsymbol{T}) \geq M_{min}$.

Построим график процентного изменения суммарного $M(\boldsymbol{T})$ от ограничения M_{min} (чтобы понять порядком каких процентных изменений мы располагаем) (Рис. 15.).

Так как параметр λ_1 отвечает в критерии как раз за $Bid \cdot CTR$ объявления, то и суммарный доход $M(\boldsymbol{T})$ по системе увеличивается с увеличением λ_1 (Рис. 15). Однако, важно заметить, что суммарный $M(\boldsymbol{T})$ растёт сильнее при варьировании λ_1 от 1.5 до 2.7, дальше же, видимо, происходит некоторое «насыщение», то есть объявлений с большими $Bid \cdot CTR$ уже не остаётся среди кандидатов на показ, поэтому мы начинаем добирать всё более дешёвые объявления и суммарный $M(\boldsymbol{T})$ начинает расти медленнее.

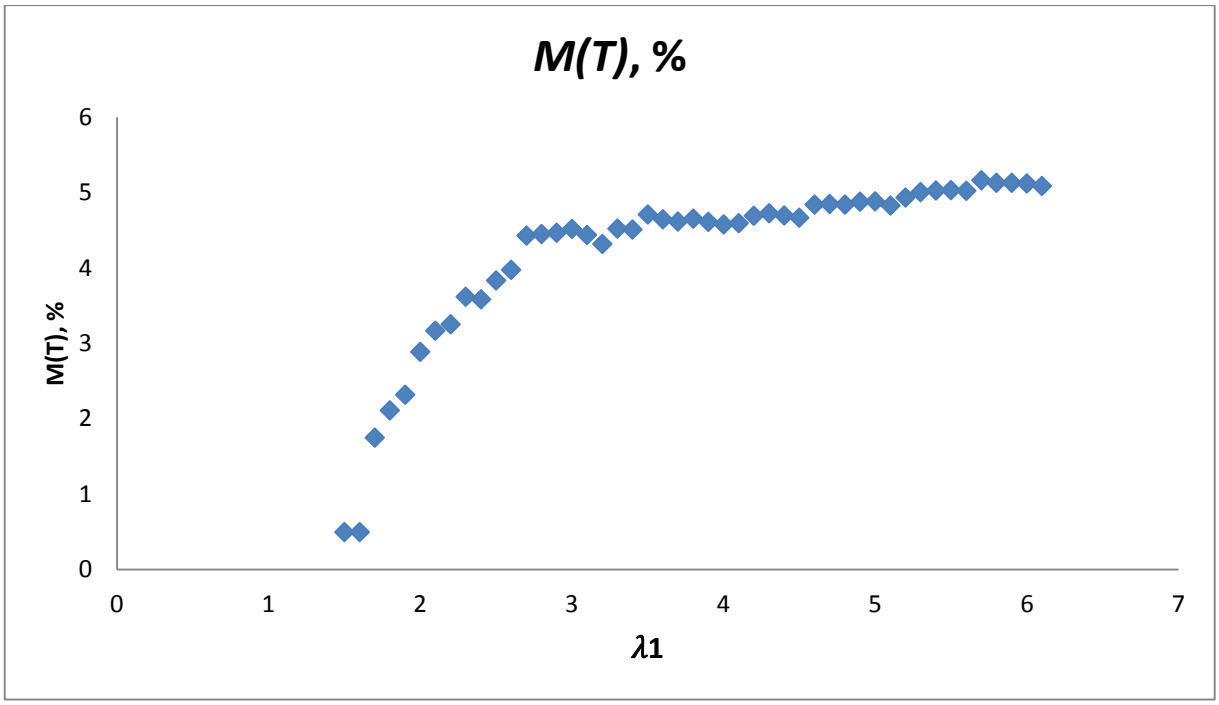

Рис. 15. Динамика изменения $M(\boldsymbol{T})$ в зависимости от λ_1.

3) *CTR vs M(T).*

Величины CTR и $M(\boldsymbol{T})$ рассчитываются следующим образом:

$$CTR = \sum_{(a,q)} t_{aq} \cdot CTR_{aq} \,/Events$$

$$M(\boldsymbol{T}) = \sum_{(a,q)} t_{aq} \cdot CTR_{aq} \cdot Bid_{aq}$$

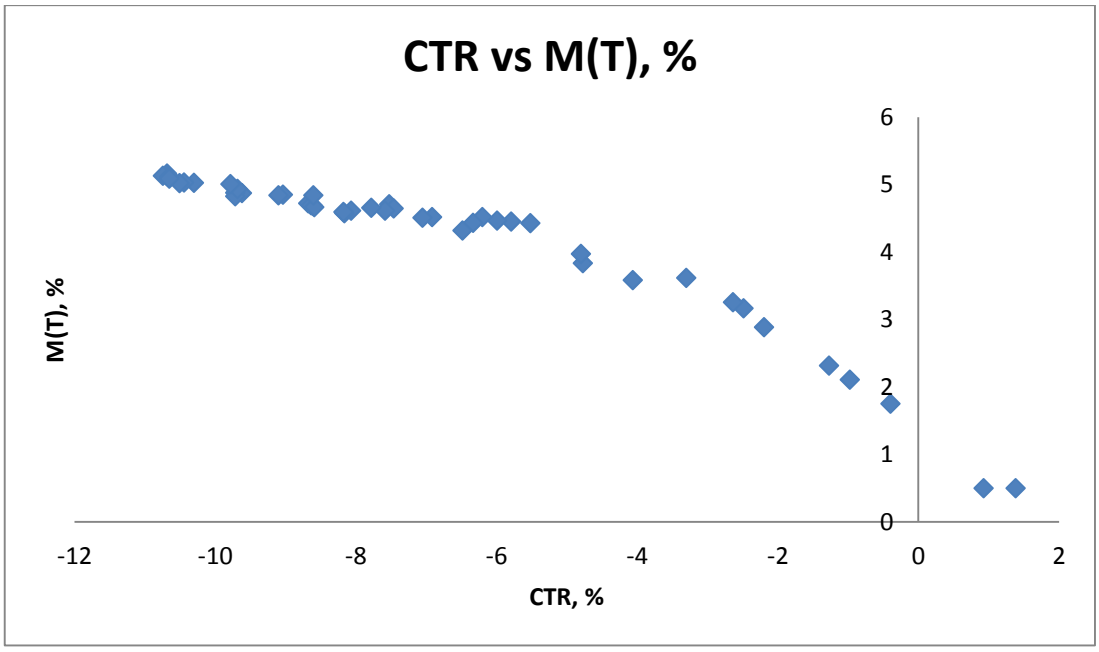

Рис. 16. $\boldsymbol{CTR\ vs\ M(T)}$ при изменении $\boldsymbol{\lambda_1}$ при фиксированном $\boldsymbol{\lambda_2}$.

По построению оптимизационной задачи важно посмотреть на зависимость суммарного показателя прогноза заработанных денежных средств от средней кликабельности (Рис. 16.).

Тут просматривается вполне логичная ситуация: при увеличении процента отличия в CTR, $M(\boldsymbol{T})$ уменьшается, то есть происходит замещение процентов одной характеристики системы на другой. Также на кривой можно выделить несколько областей: процент разницы CTR меняется от -11 до -4%: на данном участе происходит достаточно резкое увеличение CTR и совсем небольшое уменьшение процента $M(\boldsymbol{T})$: тут размен $CTR/M(\boldsymbol{T})$ достаточно большой. Далее следует отрезок изменения процентов разницы CTR от -4 до +1.4% где размен 2:1.

4) $\boldsymbol{\underline{Fillability = Events/Hits}}$ в зависимости от λ_1.

Общее количество показов рекламы в рекламном блоке над результатами поиска вычисляется следующим образом:

$$Events = \sum_{(a,q)} \boldsymbol{t}_{aq}$$

Общее количество запросов с рекламой может быть получено по следующей формуле:

$$Hits = \sum\nolimits_{(q)} \mathbb{I}\left\{\sum\nolimits_{(a)} t_{aq} > 0\right\}$$

Так как количество запросов с рекламой над результатами поиска ограничено, то величина $Fillability$ показывает заполняемость показами одного рекламного блока, то есть, сколько в среднем объявлений показалось на запрос над результатами поиска (Рис. 17.).

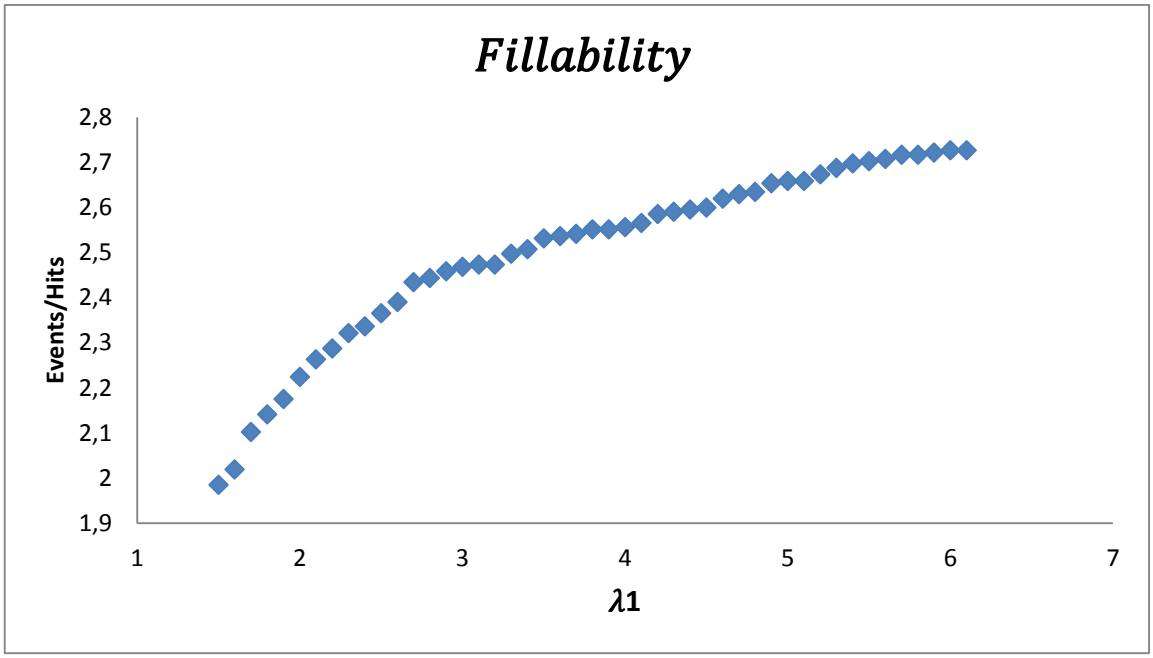

Рис. 17. Заполняемость рекламного блока над результатами поиска.

С ростом λ_1 заполняемость увеличивается, это и понятно: для того чтобы увеличить доход, нужно показывать больше объявлений. График начинается со среднего показателя заполняемости в 2 объявления в рекламном блоке на запрос, далее этот показатель увеличивается до 2.7. Важно понимать, что не всегда есть даже возможность полностью заполнить рекламный блок объявлениями, так как кандидатов на показ может быть и меньше чем три.

5) λ_3 в зависимости от λ_1.

Для фиксированного значения λ_2 с ростом λ_1 мы начинаем выбирать для показа в рекламном блоке над результатами поиска всё больше и больше объявлений (п. 2.6.3), но мы также ограничены в количестве запросов, по которым реклама может быть показана. Для того чтобы регулировать показ каждого рекламного блока, был введён параметр λ_3, который отвечает за общее качество всех рекламных объявлений, которые были выбраны для показа над результатами поиска по конкретному запросу. Интересна зависимость параметра отбора блока для показа от параметра, который отвечает за ожидаемый доход (в этом и заключается некоторая «дилемма» ограничения по количеству запросов с рекламой и ограничения по суммарному доходу) (Рис. 18.).

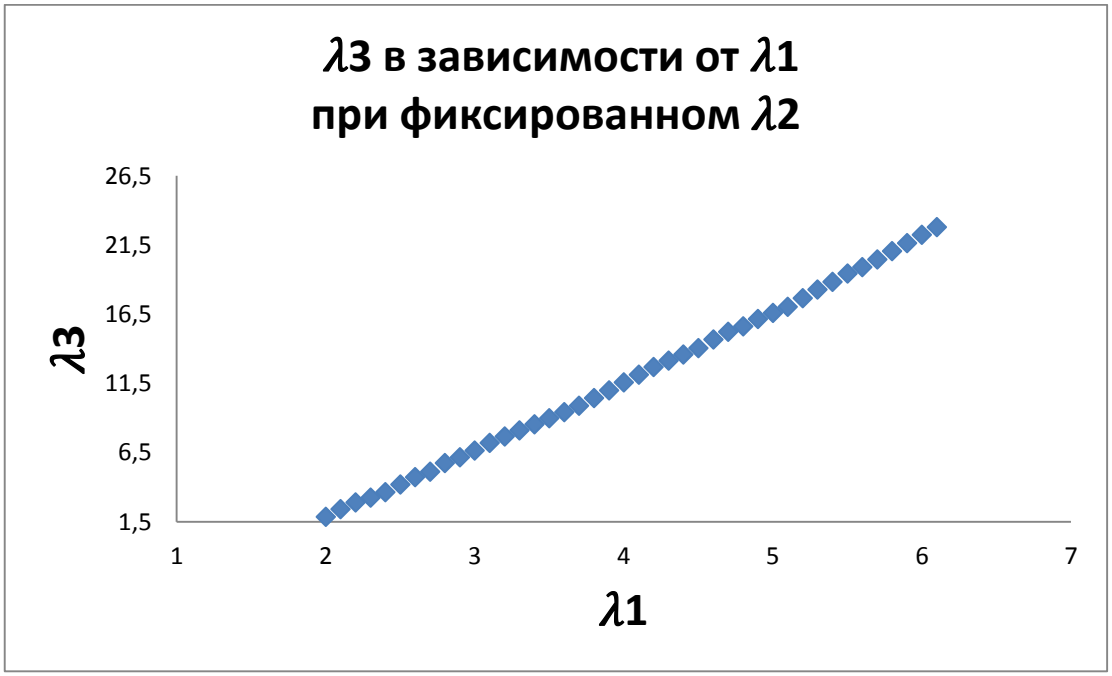

Рис. 18. λ_3 в зависимости от **λ_1** при фиксированном **λ_2**.

Видно, что при минимальных значений λ_1 ограничения на суммарный критерий показа всего рекламного блока λ_3 практически нет, но постепенно мы начинаем увеличивать λ_1 и допускать всё больше рекламных объявлений к показу, тем самым λ_3 увеличивается (для поддержания суммарного количества запросов с рекламой, т.е.

покрытия). Видно, что зависимость линейная (это связано с линейным видом функции оптимизации и ограничений, которых мы и добивались).

2.7.5 Результаты работы алгоритма: подбор всех параметров λ_1, λ_2 и λ_1.

Рассмотрим, как итеративно работает алгоритм на целом наборе запросов: для каждой пары (λ_1, λ_2) мы посчитали средний CTR и суммарный доход $M(\boldsymbol{T})$. Далее отбираются только те пары (λ_1, λ_2), где условия $Hit(\boldsymbol{T}) = \sum_{(q)} \mathbb{I}\{\sum_{(a)} t_{aq} > 0\} \leq Hit_{max}$ и $M(\boldsymbol{T}) \geq M_{min}$ выполняются (Рис. 19. 19.).

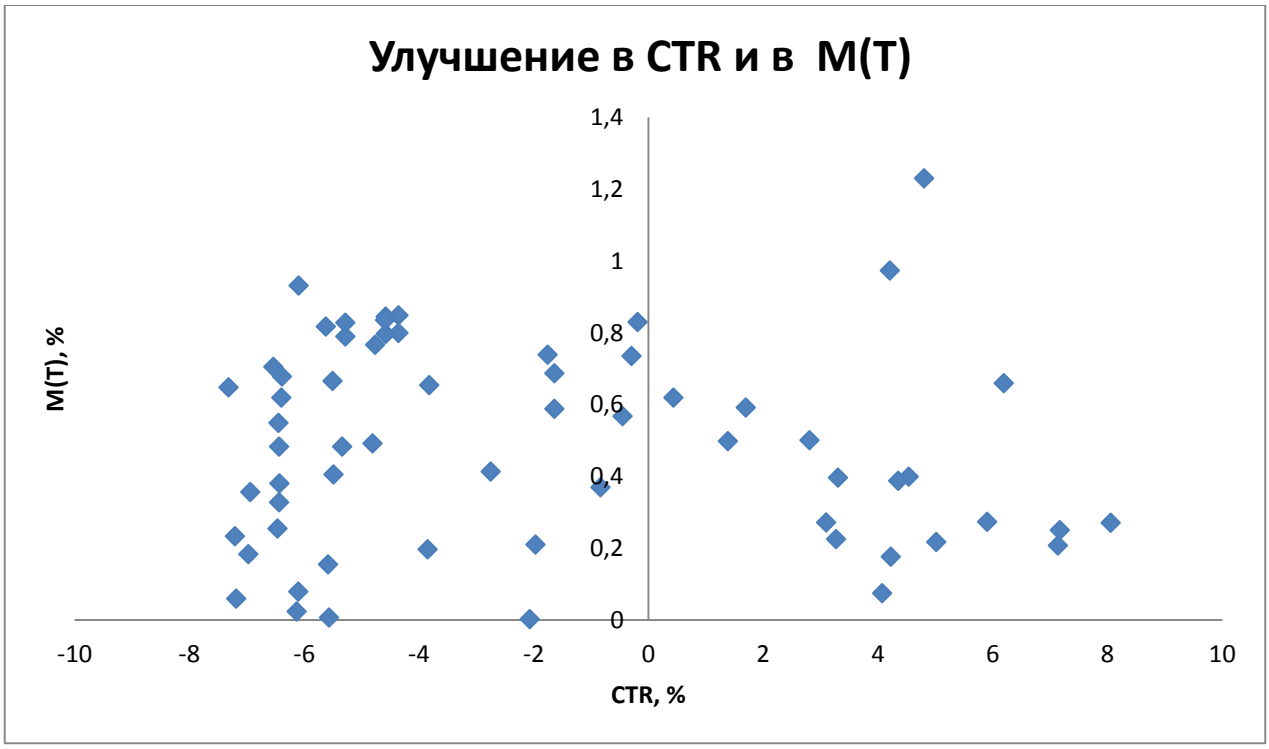

Рис. 19. Варьирование всех параметров критерия показа: лучшие точки.

По графику видно, что при выполнении всех условий оптимизации (а именно: ограничение на покрытие и ограничение по общему доходу кампании) средний CTR может отклоняться как в положительную, так и в отрицательную сторону.

Когда для каждого λ_2 выбрано оптимальное значение λ_1, при котором выполняется условие на общий доход, и λ_3, контролирующее покрытие, то алгоритм отбирает оптимальную точку по критерию $CRIT_0(\boldsymbol{T}) = \sum_{(a,q)} CTR_{aq} \cdot t_{aq}/Events$, зависимость которого от λ_2 и построим (Рис. 20.):

Каждая точка на графике (Рис.20.) представляет собой внутренний цикл, представленный на схеме алгоритма (Рис.2.). То есть при фиксированном λ_2 находится $\lambda_{1\text{опт}}(\lambda_2)$ при котором достигается ограничение на суммарный доход $M(\boldsymbol{T})$.

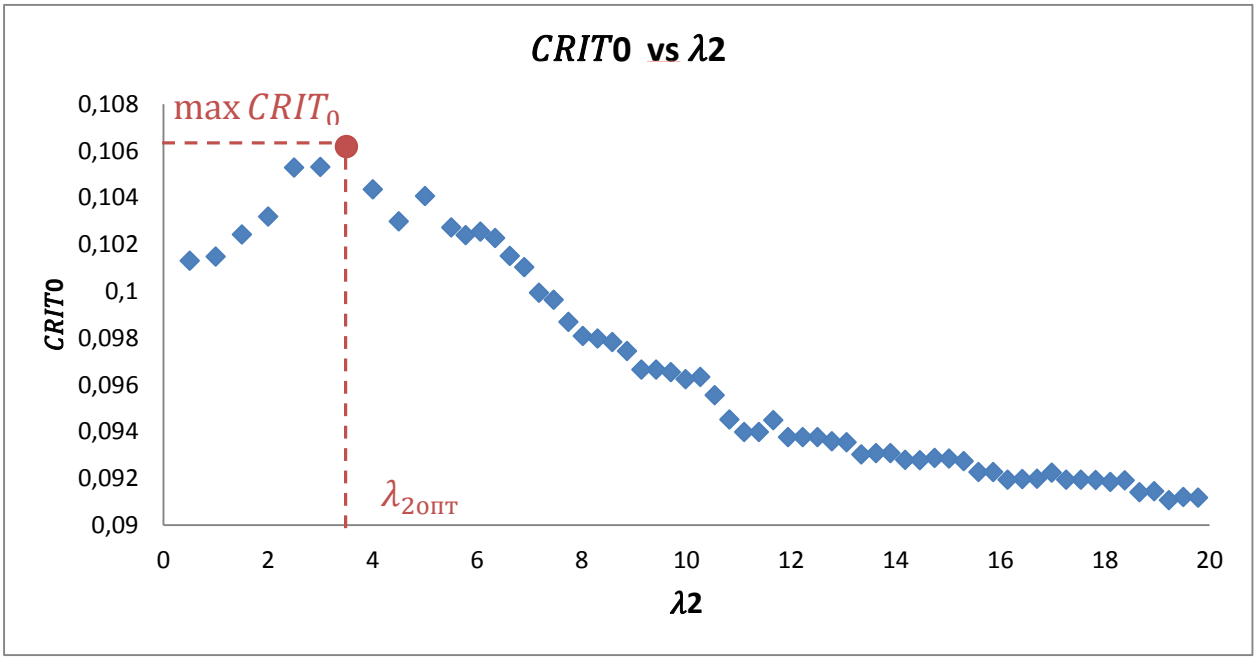

Рис. 20. Динамика оптимизационного критерия $\boldsymbol{CRIT_0}$ в зависимости от варьирования $\boldsymbol{\lambda_2}$. Выбор оптимального значения.

После того, как $\lambda_{1\text{опт}}(\lambda_2)$ найдено, мы имеем для всего набора запросов следующие характеристики:

- Для каждого объявления мы знаем значение индикатора показа t_{aq}, то есть мы знаем какие из рекламных объявлений будут показаны при выбранных значений λ_2, $\lambda_{1\text{опт}}(\lambda_2)$ и $\lambda_{3\text{опт}}(\lambda_2)$.

- Общее количество показов в рекламном блоке над результатами поиска $Events = \sum_{(a,q)} t_{aq}$
- Суммарный CTR для объявлений, показанных над результатами поиска $\sum_{(a,q)} CTR_{aq} \cdot t_{aq}$

Зная все вышеперечисленные характеристики, мы можем вычислить оптимизируемый критерий:

$$CRIT_0(\boldsymbol{T}) \quad = \sum_{(a,q)} CTR_{aq} \cdot t_{aq}/Events$$

Видно, что на графике (Рис. 20.) достигается максимум $CRIT_0(\boldsymbol{T})$, где алгоритм и остановится (дальнейшие точки предоставлены для понимания работы алгоритма). Как только максимум найден, мы знаем значение $\lambda_{2\text{опт}}$, на котором он был достигнут. После завершения работы алгоритма мы знаем оптимальные значения всех параметров критерия показа $\lambda_{1\text{опт}}$, $\lambda_{2\text{опт}}$ и $\lambda_{3\text{опт}}$ для дальнейшей работы с новыми запросами и объявлениями-кандидатами на показ в рекламном блоке над результатами поиска.

При тестировании алгоритма на реальных данных был получен off-line прирост средней кликабельности на 8%.

ГЛАВА 3. ОБОБЩЕНИЕ АЛГОРИТМА ПОДБОРА ПАРАМЕТРОВ КРИТЕРИЯ ПОКАЗА НА ПРЕДСКАЗАННУЮ ВЕРОЯТНОСТЬ КЛИКА, ЗАВИСЯЩУЮ ОТ ПОЗИЦИИ, НА КОТОРУЮ ПОПАДЁТ РЕКЛАМНОЕ ОБЪЯВЛЕНИЕ.

3.1 Основные положения по учёту позиционного эффекта.

При показе рекламного блока над результатами поиска по запросу пользователя объявления показываются одно под другим – в виде списка. Место, на котором показывается рекламное объявление, называется его **позицией**. Если в рекламном блоке было показано три объявления, то оно содержит три позиции. Первой называют самую верхнюю позицию, остальные соответственно второй и третьей (Рис.21.).

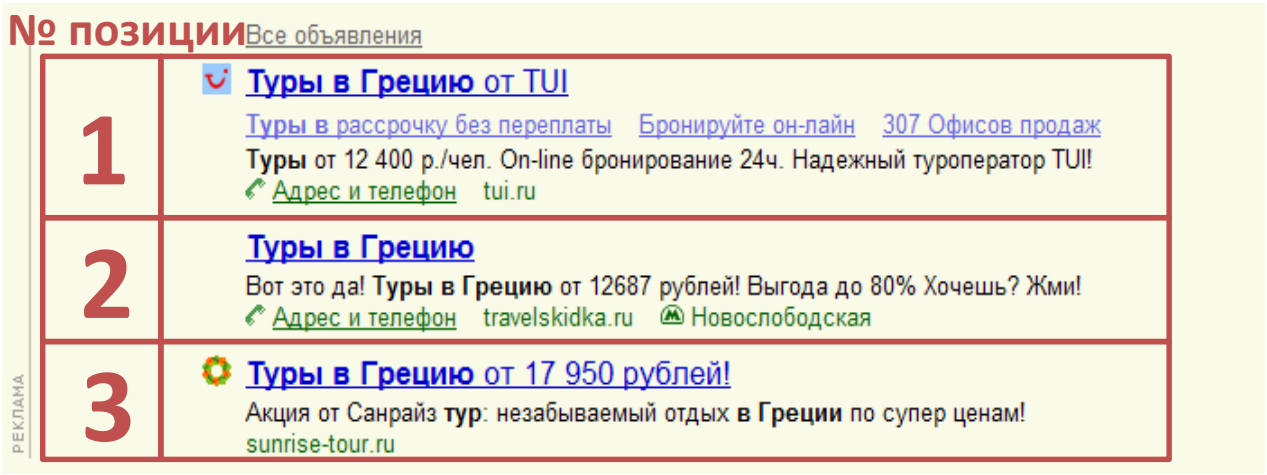

Рис. 21. Номер позиции для рекламного блока, состоящего из трёх объявлений.

Вероятность клика по объявлению зависит от того, на какой позиции оно было размещено [23]. Эксперименты по слежению за направлением взгляда пользователя при просмотре рекламных объявлений показали, что, в основном, просмотр происходит сверху в низ, то есть от первой позиции к третьей. Так же важно сказать, что пользователь больше обращает внимание на первые результаты выдачи (то есть его взгляд задерживается дольше), далее же его внимание

рассеивается [42]. Также есть ряд работ, которые показывают, что порядок просмотра пользователем рекламных объявлений предполагается не обязательно совпадающим с порядком показанных объявлений [64]. Объявления ранжируются по $w_{pos} \cdot CTR_{pos}$, где а w_{pos} – некоторый штрафной коэффициент позиции: общая идея состоит в том, что чем ниже объявление находится, тем более качественным оно должно быть, чтобы его посмотрели перед объявлением выше, но все-таки есть возможность того, что его посмотрят раньше, чем объявление выше.

Так же в статье Ксина [64] показано, что кликабельность объявления зависит от состава рекламного блока, то есть какое количество объявлений показано совместно с данным и насколько эти объявления качественны. Если вместе с некоторым объявлением A показано еще одно объявление B очень высокого качества, то внимание пользователя будет смещено в сторону объявления B, и CTR_A понизится. Если рядом с A показано объявление среднего качества, то у A будет его средний CTR. А если рядом с A показано объявление C очень низкого качества, то у A не вырастет CTR, как можно было бы ожидать из первого предложения, а упадет. Это происходит потому что вместо того, чтобы получить больше внимания от пользователя на себя, т.к. соседнее объявление – плохое, пользователь вообще разочаруется во всём рекламном блоке, так как ему показаны объявления низкого качества, и не будет кликать ни на одно из них.

В диссертационном исследовании рассматривается модель учёта позиционного эффекта размещения рекламных объявлений, которая основывается на следующих утверждениях:

- Пользователь просматривает рекламные объявления сверху вниз, и чем выше в рекламном блоке находится объявление, тем вероятнее пользователь кликает на него [68].

- В статье Фенга [31] эффект последовательного просмотра объявлений в рекламном блоке моделируется введением экспоненциального затухания. Пронумеруем сверху вниз все позиции в рекламном блоке (т.е. самое верхнее объявление имеет индекс 1). В случае, если некоторое объявление расположено на p-ой позиции в рекламном блоке, его не зависящий от позиции CTR (так называемый «истинный» CTR рекламного объявления) должен быть умножен на $\delta^{p-1}, 0 < \delta < 1$. Множитель δ определяет «затухание» внимания пользователя.
- В работе Пина [51] вводится предположение о том, что CTR можно представить в виде произведения двух сомножителей, один из которых определяется самим баннером CTR_a, а второй – позицией CTR_{pos}, на которой этот баннер находится:

$$CTR_{(a,pos)} = CTR_a \cdot CTR_{pos}$$

В работе мы будем пользоваться этим предположением при создании модельных данных: CTR_a будет зависеть только от пары запрос-объявление, а CTR_{pos} – от позиционных эффектов (в частности, в CTR_{pos} входит ранее упомянутое экспоненциальное затухание).

3.2 Математическая постановка задачи: введение позиционного эффекта.

Для решения задачи оптимизации (5) с учётом позиционного эффекта введём дополнительные обозначения и ограничения.

3.2.1 Общие обозначения.

Будем следовать системе обозначений, введенной в п. 2.1:

$\boldsymbol{P}$ – максимальное число рекламных объявлений, которое допустимо показывать в рекламном блоке в ответ на запрос пользователя (размер рекламного блока). На данный момент $P = 3$.

$\boldsymbol{p}$ – число объявлений, одновременно показанных в рекламном блоке в ответ на некоторый запрос пользователя ($1 \leq p \leq P$);

$\boldsymbol{k}$ – индекс позиции в рекламном блоке, $k \geq 1$ (самая верхняя позиция имеет индекс 1, вторая сверху – индекс 2 и так далее, Рис. 21). В дальнейшем будем придерживаться следующего соглашения: во всех величинах и выражениях, в которых одновременно присутствуют индексы k и p, эти индексы согласованы: $k \leq p$, то есть индекс позиции не может быть больше количества показанных объявлений в рекламном блоке.

$\boldsymbol{CTR_{aqpk}}$ – оценка вероятности клика при размещении a-го объявления на k-ой позиции рекламного блока в ответ на q-ый запрос пользователя, при условии, что общее число объявлений, показанных в рекламном блоке над результатами поиска в ответ на этот запрос, равно p. В случае, если такой показ невозможен (a-ое объявление несовместимо с q-ым запросом), формально полагаем эту величину равной нулю.

Считаем, что величины P и CTR_{aqpk} нам заданы, при условии того, что мы знаем состав рекламного блока: количество объявлений, которое будет показано, а так же их расстановка по позициям.

По аналогии с алгоритмом без позиционных эффектов введем бинарные переменные t_{aqpk}, такие что:

$t_{aqpk} = 1$ означает, что a–ое объявление показано над результатом поиска на q–тый запрос на k–ой позиции, при условии, что всего на этот запрос показано p объявлений в рекламном блоке;

$t_{aqpk} = 0$ в противном случае.

Кроме того, будем использовать вспомогательное обозначение:

$$Events(\boldsymbol{T}) = \sum_{(a,q,p,k)} t_{aqpk} \quad (8)$$

$Events$ – число переменных из набора $\boldsymbol{T}$, имеющих единичные значения, то есть общее число показов.

Чтобы значения, принимаемые t_{aqpk}, не противоречили друг другу (например, чтобы не получалось так, что сразу несколько разных объявлений одновременно показываются на одной и той же позиции), нужно ввести ограничения на переменные $\boldsymbol{T}$.

3.2.2 Введение ограничений, связанных с позиционностью.

Первое ограничение состоит в том, что мы не можем одновременно на одной позиции показывать два или более объявлений:

$$\forall q, p, k: \sum_a t_{aqpk} \leq 1 \tag{9}$$

Второе ограничение заключается в том, что в одном рекламном блоке не может показаться одно и тот же объявление, но на разных позициях. Если в ответ на q-ый запрос было решено показывать рекламный блок над результатами поиска, содержащее ровно p объявлений, то a-ое объявление может быть показано лишь на одной из p позиций:

$$\forall q, a, p: \sum_{k=1}^{p} t_{aqpk} \leq 1 \tag{10}$$

Третье ограничение касается согласованности индексов k и p. В силу ограничения (9) и ранее введенных соглашений об индексации $(1 \leq k \leq p \leq P)$, нельзя так выбрать значения t_{aqpk}, чтобы они обозначали ситуацию, в которой для некоторого q-го запроса было решено показывать рекламный блок, содержащий ровно p объявлений, но в действительности в нем показывается большее число объявлений (это бы означало, что $k > p$, то есть показалось всего два объявления, однако существует t_{aq23}, а это невозможно). Но все еще возможна ситуация, которая заключается в том, что для некоторого q-го запроса решено показывать ровно p объявлений, но сумма значений соответствующих t_{aqpk}, то есть сумма показов, меньшее p: $0 < \sum_{k=1}^{p} \sum_a t_{aqpk} < p$. То есть, например, $p = 3$, и объявление с

индексом a_1 должно быть показано на 1-ой позиции ($t_{a_1 q 31} = 1$), объявление с индексом a_3 – на 3-ей позиции ($t_{a_3 q 33} = 1$). Но нет ни одного объявления, которое могло бы показаться на 2-ой позиции: не существует такого индекса a_2, чтобы было выполнено $t_{a_2 q 32} = 1$.

Введем условие, которое сделает подобные ситуации невозможными:

$$\forall q, p\colon \sum_{k=1}^{p} \sum_{a} t_{aqpk} \in \{0, p\} \qquad (11)$$

3.2.3 Введение ограничений, связанных с показом рекламного блока.

Определим еще один набор бинарных переменных:

$\tau_{qp} = 1$, если в ответ на q-ый запрос показан рекламный блок, содержащее ровно p рекламных объявлений $\sum_{k=1}^{p} \sum_{a} t_{aqpk} = p$;

$\tau_{qp} = 0$ иначе.

Ограничение состоит в том, что если для q-го запроса решено показывать рекламный блок над результатами поиска, то число содержащихся в нем объявлений должно быть однозначно определено. Необходимо исключить такие события, когда показывается ровно два объявления и в то же время ровно три объявления. Это ограничение можно записать так:

$$\forall q\colon \sum_{p=1}^{P} \tau_{qp} \leq 1 \qquad (12)$$

Таким образом, сумма $\sum_{p=1}^{P} \tau_{qp} = 1$, если для q-го запроса решено показать рекламный блок над результатами поиска, и $\sum_{p=1}^{P} \tau_{qp} = 0$ иначе.

3.2.4 Ограничения из базовой задачи оптимизации в случае учёта позиционного эффекта. Математическая постановка задачи.

С учётом обозначений, введённых выше, ограничение на покрытие (3) примет вид:

$$\sum_q \sum_{p=1}^{P} \tau_{qp} \le H_{max} \quad (13)$$

Если выполнены ограничения (9) – (12), то значения индикаторов $\boldsymbol{T}$ не противоречат друг другу. Целевая функция средней кликабельности (1) запишется в следующем виде:

$$\overline{CTR}(\boldsymbol{T}) = \sum_{(a,q,p,k)} CTR_{aqpk} \cdot t_{aqpk} \,/ Events(\boldsymbol{T}) \quad (14)$$

То есть просуммируем значения кликабельностей для всех показанных объявлений в зависимости от их позиций в рекламном блоке и разделим на суммарное количество показанных объявлений по всем запросам.

Ожидаемый доход запишется в виде:

$$M(\boldsymbol{T}) = \sum_{(a,q,p,k)} CTR_{aqpk} \cdot Bid_a \cdot t_{aqpk} \quad (15)$$

По условию изначальной оптимизационной задачи из п. 2.1 ожидаемый доход $M(\boldsymbol{T})$ должен быть не меньше заданной неотрицательной величины M_{min}:

$$M(\boldsymbol{T}) \ge M_{min} \quad (16)$$

Будем решать задачу максимизации средней кликабельности по переменным t_{aqpk}:

$$CRIT_0(\boldsymbol{T}) = \overline{CTR}(\boldsymbol{T}) \to \max_{\boldsymbol{T}} \quad (17)$$

при условиях (9) – (13), (16).

3.3 Решение задачи оптимизации с учётом позиционного эффекта.

3.3.1 Расширение области значений переменных t_{aqpk}.

Как и в п. 2.2.1 позволим переменным t_{aqpk} принимать любые значения из отрезка $[0;1]$, а не только крайние значения 0 и 1. Отсюда вытекает еще одно ограничение-неравенство:

$$\forall a, q, p, k:\ 0 \le t_{aqpk} \le 1 \qquad (18)$$

Результаты будут такими же, как если бы мы решали задачу без расширения области значений t_{aqpk}: окажется, что для максимизации критерия (17) переменные t_{aqpk} все равно должны принимать только крайние значения 0 и 1.

3.3.2 Перевод ограничения по суммарному доходу в критерий

Прежде всего, переведем ограничение (16) в критерий при помощи множителя Лагранжа $\lambda_1 \ge 0$ и будем искать максимум получившегося критерия при оставшихся ограничениях (9) – (13). В результате получаем новый критерий:

$$CRIT_1(\boldsymbol{T}, \lambda_1) = \overline{CTR}(\boldsymbol{T}) - \lambda_1(M_{min} - M(\boldsymbol{T})) \qquad (19)$$

3.3.3 Перевод ограничения по суммарным денежным средствам в критерий

Как и в п. 2.2 было бы удобно оптимизировать критерий $CRIT_1$, если бы он представлял собой сумму слагаемых. Введем дополнительное ограничение-равенство:

$$Events(\boldsymbol{T}) = E_0, E_0 > 0 \qquad (20)$$

Это ограничение переведем в критерий при помощи множителя Лагранжа λ_1:

$$CRIT_2(\boldsymbol{T}, \lambda_1, \lambda_2, E_0) = \overline{CTR}(\boldsymbol{T}) - \lambda_1\big(M_{min} - M(\boldsymbol{T})\big) - \lambda_2(Events(\boldsymbol{T}) - E_0) \qquad (21)$$

3.3.4 Декомпозиция задачи

Для $CRIT_2$ мы уже можем провести декомпозицию: представить его в виде суммы слагаемых, каждое из которых зависит только от

одной из переменных t_{aqpk}. Преобразуем $CRIT_2$, сгруппировав слагаемые, зависящие от t_{aqpk}:

$$CRIT_2(\boldsymbol{T}, \lambda_1, \lambda_2, E_0) = \frac{\sum_{(a,q,p,k)} CTR_{aqpk} \cdot t_{aqpk}}{E_0} - \lambda_1 \big(M_{min} - \sum_{(a,q,p,k)} CTR_{aqpk} \cdot Bid_a \cdot t_{aqpk}\big) - \lambda_2 \big(\sum_{(a,q,p,k)} t_{aqpk} - E_0\big) = \sum_{(a,q,p,k)} \left(\frac{CTR_{aqpk}}{E_0} + \lambda_1 \cdot CTR_{aqpk} \cdot Bid_a - \lambda_2\right) \cdot t_{aqpk} + (-\lambda_1 \cdot M_{min} - E_0) = \sum_q \sum_{(a,p,k)} F_{aqpk}(\lambda_1, \lambda_2, E_0) \cdot t_{aqpk} + C = \sum_q R_q(\boldsymbol{T}, \lambda_1, \lambda_2, E_0) + C$$

В результате получилось представить $CRIT_2$ в виде суммы слагаемых, каждое из которых относится только к одному запросу, и константы, не зависящей от переменных $\boldsymbol{T}$. Для максимизации критерия $CRIT_2(\boldsymbol{T}, \lambda_1, \lambda_2, E_0)$ при фиксированных λ_1, E_0 и λ_2 по переменным $\boldsymbol{T}$ и при условиях (9) – (13), (16) мы будем отдельно максимизировать по $\boldsymbol{T}$ при условиях (9) – (13), (16) каждое из слагаемых $R_q(\boldsymbol{T}, \lambda_1, \lambda_2, E_0)$ а затем отдельно учитывать ограничение (13) на покрытие.

3.3.5 Максимизация R_q с учётом ограничений.

Переменные t_{aqpk}, соответствующим образом как и в п. 2.2.3, могут принимать значения только 0 и 1. Все нужные значения CTR_{aqpk} и Bid_a нам заданы и, зная λ_1, E_0 и λ_2, можно вычислить:

$$F_{aqpk}(\lambda_1, \lambda_2, E_0) = \frac{CTR_{aqpk}}{E_0} + \lambda_1 \cdot CTR_{aqpk} \cdot Bid_a - \lambda_2.$$

Заметим, что если $F_{aqpk} \leq 0$, то соответствующий множитель t_{aqpk} должен быть равен нулю (ненулевое t_{aqpk} может быть только положительным в силу (20), поэтому в противном случае слагаемое $F_{aqpk} \cdot t_{aqpk} < 0$ будет лишь уменьшать сумму $R_q = \sum_{(a,p,k)} F_{aqpk} \cdot t_{aqpk}$). Напротив, если $F_{aqpk} > 0$, то в целях максимизации R_q это F_{aqpk} должно быть умножено на максимально значение t_{aqpk}, то есть на единицу. Так мы и будем поступать в дальнейшем.

3.3.6 Подбор оптимального числа баннеров на показ.

Ограничение (12) сводится к тому, что доля запросов, на которые можно показывать рекламный блок над результатами поиска, ограничена, при этом нам хотелось бы, чтобы число показанных рекламных объявлений, их состав и позиция в рекламном блоке были оптимальны с точки зрения максимизации соответствующего $R_q(\boldsymbol{T})$. Отдельно рассмотрим ситуации, когда показывается только одно объявление, два объявления, и так далее до P. Для каждой из этих ситуаций найдём оптимальный набор объявлений – таким образом будут найдены все оптимальные варианты показа рекламных объявлений в рекламном блоке по данному запросу. В итоге выберем тот вариант, который обеспечивает максимальное значение R_q:

$$\max_{\boldsymbol{T}} R_q(\boldsymbol{T}, \lambda_1, \lambda_2, E_0) = max\,{}_{p=1}^{P} \left(\max_{\boldsymbol{T}} \left(\sum_{(a,k)} F_{aqpk}(\lambda_1, \lambda_2, E_0) \cdot t_{aqpk} \right)\Big|_{(9)-(13),(16)} \right) \quad (22)$$

Учитывая, что в рекламном блоке над результатами поиска допускается одновременный показ не более трёх объявлений, объем такого перебора мал. Результат максимизации R_q можно представлять в виде списка из индексов рекламных объявлений. Если, допустим, этот список имеет вид (a_1, a_2), то по q-му запросу оптимально показать рекламный блок, в котором на первой позиции расположено объявление с индексом a_1, на второй – объявление с индексом a_2. Вполне может оказаться, что при заданных λ_1, λ_2 и E_0 для некоторых значений p (или даже для всех) вообще невыгодно показывать рекламный блок над результатами поиска, либо задача $\sum_{(a,k)} F_{aqpk}(\lambda_1, \lambda_2, E_0) \cdot t_{aqpk} \to \max \boldsymbol{T}$ не решается при ограничениях (9) – (13) и (16), поэтому этот список может быть пустым.

3.3.7 Отбор рекламных объявлений и их размещение при заданном числе показов.

Итак, для того чтобы решить задачу максимизации R_q нам остается рассмотреть решение задачи вида $\sum_{(a,k)} F_{aqpk}(\lambda_1, \lambda_2, E_0) \cdot t_{aqpk} \to \max \boldsymbol{T}$ при ограничениях (9) – (13) и (16). Расположим величины F_{aqpk} (при фиксированных q и p) в прямоугольной матрице с p строками и N столбцами:

$$\begin{pmatrix} F_{q1p1} & \cdots & F_{pNp1} \\ \vdots & \ddots & \vdots \\ F_{q1pp} & \cdots & F_{qNpp} \end{pmatrix} \tag{23}$$

Если некоторое t_{aqpk} (при данных фиксированных q и p) положить равным единице, то тем самым мы отберем из этой матрицы в сумму $\sum_{(a,k)} F_{aqpk} \cdot t_{aqpk}$ соответствующий элемент F_{aqpk}. Из-за ограничений (9) и (10) становится невозможным произвольное установление значения t_{aqpk} равным единице, если соответствующий элемент F_{aqpk} положителен, и нулю в противном случае.

Заметим, что условие (10) означает, что в этой матрице из любой строки можно выбрать не более одного элемента. Аналогично, условие (11) означает, что из любого столбца матрицы можно выбрать не более одного элемента. По условию задачи оптимизации необходимо, чтобы сумма выбранных элементов была как можно больше. В такой постановке мы получаем так называемую **задачу о назначениях**, которая решается **венгерским алгоритмом** [45].

При помощи венгерского алгоритма мы выбираем из каждой строки матрицы (23) по одному значению F_{aqpk} так, чтобы никаких два выбранных значения не находились в одном столбце. При этом сумма выбранных значений максимальна среди всех возможных выборок,

удовлетворяющих условиям нахождения в разных строках и разных столбцах.

Просмотрим выбранные p значений F_{aqpk}: если все они положительны, то положим соответствующие им значения t_{aqpk} равными единице, а все прочие t_{aqpk} (при фиксированных q и p) – нулю, тогда условия (9) и (10) будут выполнены. Также, очевидно, в этом случае выполнится и условие (11). То есть мы нашли решение (распределение рекламных объявлений по позициям) задачи $\sum_{(a,k)} F_{aqpk} \cdot t_{aqpk} \to \max \boldsymbol{T}$, удовлетворяющее всем ограничениям. В том случае, если какое-то из выбранных значений F_{aqpk} не положительно, то придется его не рассматривать, и, следовательно, придется исключить из рассмотрения и остальные выбранные значения, иначе не будет выполняться условие (11). В этом случае мы полагаем все t_{aqpk} равными нулю.

3.3.8 Учёт ограничения на покрытие.

Итак, мы выполнили максимизацию по переменным $\boldsymbol{T}$ критерия $CRIT_2$ при ограничениях (9) – (13), (16) и данных фиксированных значениях λ_1, λ_2 и E_0, пользуясь тем, что $CRIT_2$ разбивается на сумму M независимых величин $R_q(\boldsymbol{T}, \lambda_1, \lambda_2, E_0)$, каждая из которых соответствует одному из запросов. На данный момент для каждого запроса найден список рекламных объявлений, который представляет собой оптимальный набор для показа в рекламном блоке над результатами поиска в ответ на этот запрос. Некоторые списки могут быть пустыми: это значит, что при данных λ_1, λ_2 и E_0 по этому запросу не нашлось объявлений, удовлетворяющим условиям оптимизационной задачи, и рекламный блок показывать не стоит. Чтобы учесть ограничение (13) на покрытие для данного набора запросов, мы должны показывать рекламный блок в ответ на не более

чем H_{max} запросов из этого набора. Если число запросов с непустым списком меньше, чем H_{max}, то ограничение уже автоматически выполнено. В противном случае упорядочим списки по убыванию соответствующих им значений R_q, и оставим только первые H_{max} списков. Тогда значение $CRIT_2$ будет равно сумме значений R_q по оставленным спискам. Тем самым будет удовлетворено ограничение на покрытие, и $CRIT_2$ принимает максимально возможное при этом значение. Кроме того, необходимо вычислить еще одну величину: λ_3, которая равна минимальному значению R_q по оставленным (непустым) спискам. Эта величина зависит от $\lambda_1,\ \lambda_2$ и E_0: $\lambda_3 = \lambda_3(\lambda_1,\ \lambda_2, E_0)$. Мы перебором находим оптимальные $\lambda_{1\ \text{опт}},\ \lambda_{2\ \text{опт}}$ и $E_{0\ \text{опт}}$, при которых достигается максимум по $\boldsymbol{T}$ значения критерия $CRIT_2(\boldsymbol{T}, \lambda_1,\ \lambda_2, E_0)$ и при этом выполняются ограничения. Оптимальное значение $\lambda_{3\ \text{опт}} = \lambda_3(\lambda_{1\ \text{опт}},\ \lambda_{2\ \text{опт}}, E_{0\ \text{опт}})$ надо запомнить, так как оно будет использоваться при работе с новыми запросами для учета ограничения на покрытие.

3.3.9 Общая схема оптимизации.

Общая схема оптимизации будет иметь следующий вид:

Алгоритм можно записать в виде:

Цикл по λ_2

Величина λ_2 доставляет максимум нашему основному критерию $CRIT_0(\boldsymbol{T})$.

Цикл по λ_1

Вырученные деньги с ростом λ_1 не убывают, то есть можно сказать что λ_1– это параметр, регулирующий поступление денег от рекламодателей.

Перебор по λ_1 идёт до достижения равенства

$M(\boldsymbol{T}) = M_{min}$

Цикл по запросам q

По всем объявлениям-кандидатам a для запроса q:

Составить «список для показа», вычисляя для каждого объявления:

$$F_{aqpk}(\lambda_1, \lambda_2, E_0) = \frac{CTR_{aqpk}}{E_0} + \lambda_1 \cdot CTR_{aqpk} \cdot Bid_a - \lambda_2.$$

С помощью венгерского алгоритма решаем задачу максимизации:

$\sum_{(a,k)} F_{aqpk} \cdot t_{aqpk} \to \max \boldsymbol{T}$, в список для показа попадают соответствующие t_{aqpk}. Таким образом:

$$\boldsymbol{T} = \arg\max_{\boldsymbol{T}} \left(CRIT_2(\boldsymbol{T}, \lambda_1, \lambda_2)|_{(9)-(13),(16)} \right)$$

Вычислить вклад в суммарный критерий объявлений, вошедших в список для показа в рекламный блок над результатами поиска:

$$R_q = \sum_{(a,k)} F_{aqpk} \cdot t_{aqpk}$$

Конец цикла по запросам q

Упорядочить списки для запросов в порядке убывания R_q.

Оставить только первые H_{max} из них (выполнение ограничения по покрытию), остальные обнулить.

Запомнить $\boldsymbol{\lambda_3} = \min_q R_q$

Положить:

$$t_{aqpk} = \begin{cases} 1, & \text{если пара } (a, q) \text{ оставлена для показа,} \\ 0, & \text{в противном случае} \end{cases}$$

Менять λ_1, пока не будет достигнуто $M(\boldsymbol{T}) = M_{min}$ следующим образом:

Если $M(\boldsymbol{T}) < M_{min}$, то уменьшить λ_1

Если $M(\boldsymbol{T}) > M_{min}$, то увеличить λ_1

Положить $\boldsymbol{\lambda_{1\,\text{опт}}} = \lambda_1$.

Запомнить λ_3, $\lambda_{1\,\text{опт}}$, списки t_{aqpk} (то есть рекламные объявления, отобранные для показа)

Конец цикла по λ_1.

Менять λ_2 чтобы достигнуть ***максимума*** $CRIT_0(\boldsymbol{T}) = \sum_{(a,q,p,k)} CTR_{aqpk} \cdot \boldsymbol{t}_{aqpk}/Events(\boldsymbol{T})$

Положить $\boldsymbol{\lambda_{2\,\text{опт}}}$ то значение λ_2, при котором достигнут $\max CRIT_0(\boldsymbol{T})$

Конец цикла по λ_2

В конце работы алгоритма мы получаем величины $\lambda_{1\,\text{опт}}$, $\lambda_{2\,\text{опт}}$ и $\lambda_{3\,\text{опт}}$, которые будут использоваться для работы с новыми запросами.

3.3.10 Схема работы с новыми запросами.

В результате работы алгоритма на обучающем наборе запросов получаются значения параметров $\lambda_{1\,\text{опт}}$, $\lambda_{2\,\text{опт}}$ и $\lambda_{3\,\text{опт}}$. Будем использовать эти значения при работе с новыми запросами. Считаем, что для новых запросов и соответствующих им рекламных объявлений известны величины CTR_{aqpk} и Bid_a, либо их можно каким-то образом получить. Тогда алгоритм обработки новых запросов выглядит следующим образом (для -го нового запроса):

Итак, получив новый запрос нужно:

1) Отобрать объявления-кандидаты для возможных показов (по фразам запроса).
2) Для каждого из этих объявлений известно значение ставки Bid_a и прогноз кликабельности CTR_{aqpk}, где q – индекс запроса, a – индекс баннера, p – индекс позиции объявления среди показанных, k – общее количество рекламных объявлений в рекламном блоке.
3) для каждого возможного размера p рекламного блока:

3.1. вычислить значения $F_{aqpk} = CTR_{aqpk} + \lambda_{1\,\text{опт}} \cdot CTR_{aqpk} \cdot Bid_a - \lambda_{2\,\text{опт}}$

3.2. решить задачу $\sum_{(a,k)} F_{aqpk} \cdot t_{aqpk} \rightarrow \max \boldsymbol{T}$ при ограничениях (9) – (12)

3.3. выбрать наилучший вариант по p с точки зрения максимизации:

$$R_q = \sum_{(a,p,k)} F_{aqpk} \cdot t_{aqpk}$$

4) Если величина R_q меньше $\lambda_{3\,\text{опт}}$, то обнулить список, и для данного запроса не показывать рекламный блок над результатами поиска. Иначе показать в рекламном блоке все объявления, которые отобрались для показа в процессе п. 3.2.

После проведения операций 1) - 4), будет совершенно ясно показывать ли по новому поступившему запросу объявления, если да, то какое количество рекламных объявлений показывать, конкретно какие из всех объявлений-кандидатов на показ и в какой последовательности.

3.4 Численный эксперимент на модельных данных.

3.4.1 Создание модельных данных.

Для отладки описанного алгоритма отбора объявлений с учетом позиционных эффектов, его тестирования и сравнения с аналогичным алгоритмом, не учитывающим позиционные эффекты, будем использовать модельные данные, из которых и будет состоять обучающий набор запросов.

В статье Пина [51] предполагалось моделировать ставки логнормальным распределение с параметрами $\mu = 1$ и $\sigma = 0.1$. Мы будем пользоваться этим предположением при генерации модельных

данных, за исключением того, что параметры могут несколько отличаться от указанных там $\mu = 1$ и $\sigma = 0.1$.

Для генерации значений CTR считаем, что значения CTR имеют бета-распределение с некоторыми параметрами. Предположение о том, что значения кликабельности имеют бета-распределение, является часто используемым [63], и оно хорошо подтверждается на практике.

Когда выше утверждалось, что значения CTR можно представлять как реализацию некоторой случайной величины, имеющей бета-распределение, имелись в виду значения CTR без учета позиционных эффектов: CTR(*запрос, объявление*). Далее такие кликабельности будем также обозначать как CTR_{aq}: оценка вероятности клика на a-ое объявление в случае, если оно показывается на странице результатов поиска в ответ на q-ый запрос. Для того чтобы перейти от вероятностей CTR_{aq} кликов без учета позиционных эффектов к вероятностям кликов CTR_{aqpk}, зависящим от позиционных эффектов, используем предположение, в котором вероятность клика CTR представляется в виде произведения:

$$CTR = CTR_{aq} \cdot CTR_{pos}$$

Множитель CTR_{aq} зависит только от объявления и запроса, а множитель CTR_{pos} вводится для учета позиционных эффектов. Такое предположение рассматривается во многих работах [68], [31], [47]. В нашем случае множитель CTR_{pos} представляется в виде $CTR_{pos} = \gamma_p \cdot decay_q^{k-1}$. Множитель γ_p учитывает тот факт, что любое объявление в рекламном блоке может быть показано одно, либо иметь одного, либо двух соседей (напомним, что всего в рекламном блоке допускается одновременный показ не более трёх рекламных объявлений). То есть если в ответ на q-ый запрос в рекламном блоке над результатами поиска показываются два объявления, в число которых входит a-ое, то

для учета этого факта в оценке вероятности клика для a-го объявления производится умножением CTR_{aq} на γ_2. Значения для γ_1, γ_2 и γ_3 выбираются из следующих предположений. Когда объявление показывается в одиночку, вероятность клика на него гораздо больше, чем когда он показывается совместно с одним или более рекламных объявлений. Фактически значение кликабельности CTR_{aq11} (объявление показывается в одиночестве), должно немного превосходить значение кликабельности CTR_{aq} (итоговое усредненное значение без позиционных эффектов), поэтому будем полагать $\gamma_1 \approx 1$, $\gamma_1 > 1$. Далее, когда объявление показывается в компании ровно одного соседа, то вероятность того, что пользователь кликнет на него, больше, чем когда это же объявление показывается в компании ровно двух соседей. Поэтому будем полагать, что $1 \geq \gamma_2 > \gamma_3$. Например, в нескольких вычислительных экспериментах использовались значения $\gamma_1 = 1.065$, $\gamma_2 = 0.92$ и $\gamma_3 = 0.83$, а в нескольких других $\gamma_1 = 1.045$, $\gamma_2 = 0.72$ и $\gamma_3 = 0.902$.

Теперь рассмотрим множитель $decay_q^{k-1}$ (величина $decay_q$, возведенная в степень $k-1$). В работе Фенга [31] предлагалось ввести экспоненциальное затухание для того, чтобы учесть позицию объявления. Допустим, что в ответ на q-ый запрос показывается рекламный блок над результатами поиска, и в нем показано ровно три объявления, причем интересующее нас a-ое объявление находится на k-ой сверху позиции $(1 \leq k \leq 3)$. Тогда учет этого эффекта производится умножением кликабельности CTR_{aq} на $decay_q^{k-1}$. Значения множителя $decay_q$ можно взять из непрерывного распределения на отрезке [0; 1] – например, из бета-распределения с параметрами α_{decay} и β_{decay}.

Схема, по которой генерируются модельные данные, такова:

1. Задаемся:
   - параметрами $\mu_{Bid} \geq 0$ и $\sigma_{Bid} \geq 0$ для логнормального распределения ставок;
   - параметрами $\alpha_{CTR} > 0$ и $\beta_{CTR} > 0$ для бета-распределения не зависящих от позиционных эффектов кликабельностей CTR_{aq};
   - параметрами $\alpha_{decay} > 0$ и $\beta_{decay} > 0$ для бета-распределения коэффициентов затухания, которые используются для учета позиции k, на которой находится данное объявление;
   - положительными множителями γ_1, γ_2 и γ_3, используемыми для учета числа p объявлений, показываемых одновременно в рекламном блоке;
   - $\boldsymbol{M}$ – размер модельного набора запросов;
   - $\boldsymbol{N}$ – общее число различных объявлений;
   - минимальным ($\boldsymbol{ads\_min}$) и максимальным ($\boldsymbol{ads\_max}$) числом рекламных объявлений, которые могут быть кандидатами для показа в рекламном блоке по одному запросу.
2. Будем различать запросы по идентификатору q, объявления – по индексу/идентификатору a. То есть на данный момент у нас есть $\boldsymbol{M}$ запросов, индексируемых при помощи q, и $\boldsymbol{N}$ рекламных объявлений, индексируемых при помощи a. Пока что это пустые сущности: не составлены списки объявлений, соответствующих запросам, не сгенерированы ставки для запросов и т.д.
3. Для каждого запроса q сгенерируем ставку Bid_a, являющуюся реализацией случайной величины из логнормального распределения с параметрами μ_{Bid} и σ_{Bid}.
4. Для каждого запроса q сгенерируем коэффициент затухания $decay_q$ из бета-распределения с параметрами α_{decay} и β_{decay}.
5. Для каждого запроса q:

5.1. Определяем, сколько объявлений ему соответствует (равномерная случайная целая величина между ads_{min} и ads_{max}) – назовем это число $corr\_ads\_number_q$.

5.2. Делаем случайную выборку размера $corr\_ads\_number_q$ из множества идентификаторов a. Объявления с выбранными идентификаторами теперь соответствуют запросу q.

5.3. Для текущего запроса q и каждого объявления a из соответствующих ему объявлений генерируем значения CTR_{aq}.

5.4. Для каждого возможного числа баннеров p, показываемых в рекламном блоке над результатами поиска определяем CTR_{aqpk} как $CTR_{aqpk} = CTR_{aq} \cdot \gamma_p \cdot decay_q^{k-1}$. Произведение $\gamma_p \cdot decay_q^{k-1}$ моделирует CTR_{pos}.

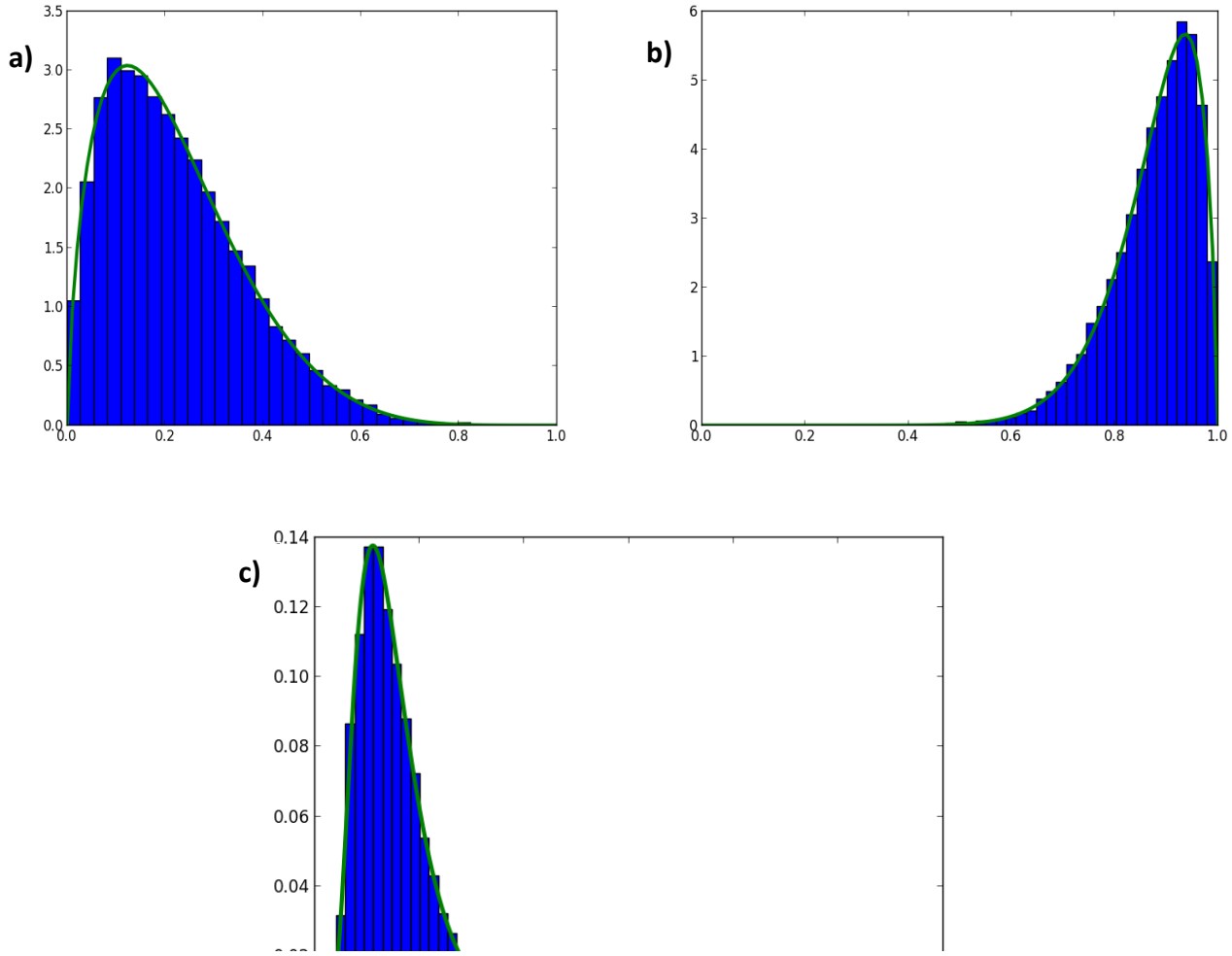

Рис. 22. Гистограммы распределений и плотности для ***CTR*** (a), коэффициентов затухания (b) и ставок (c).

Примеры гистограмм и функций плотности распределения для позиционно-независимых CTR_{aq}, коэффициентов затухания $decay_q$ и ставок Bid_a приведены на Рис. 22-23. Гистограммы (Рис. 22.) были получены при следующих значениях параметров: $\alpha_{CTR} = 1.7,\ \beta_{CTR} = 6,\ \alpha_{decay} = 12.8,\ \beta_{decay} = 1.8,\ \mu_{Bid} = 1.93,\ \sigma_{Bid} = 0.47,\ M = 1000,\ N = 12400,\ ads_{min} = 5,\ ads_{max} = 40.$

Гистограммы (Рис. 23) были получены при таких значениях параметров:$\alpha_{CTR} = 0.5,\ \beta_{CTR} = 12.5,\ \alpha_{decay} = 17.8,\ \beta_{decay} = 0.9,\ \mu_{Bid} = 4.13,\ \sigma_{Bid} = 0.67,\ M = 1400,\ N = 500000,\ ads_{min} = 5,\ ads_{max} = 40.$

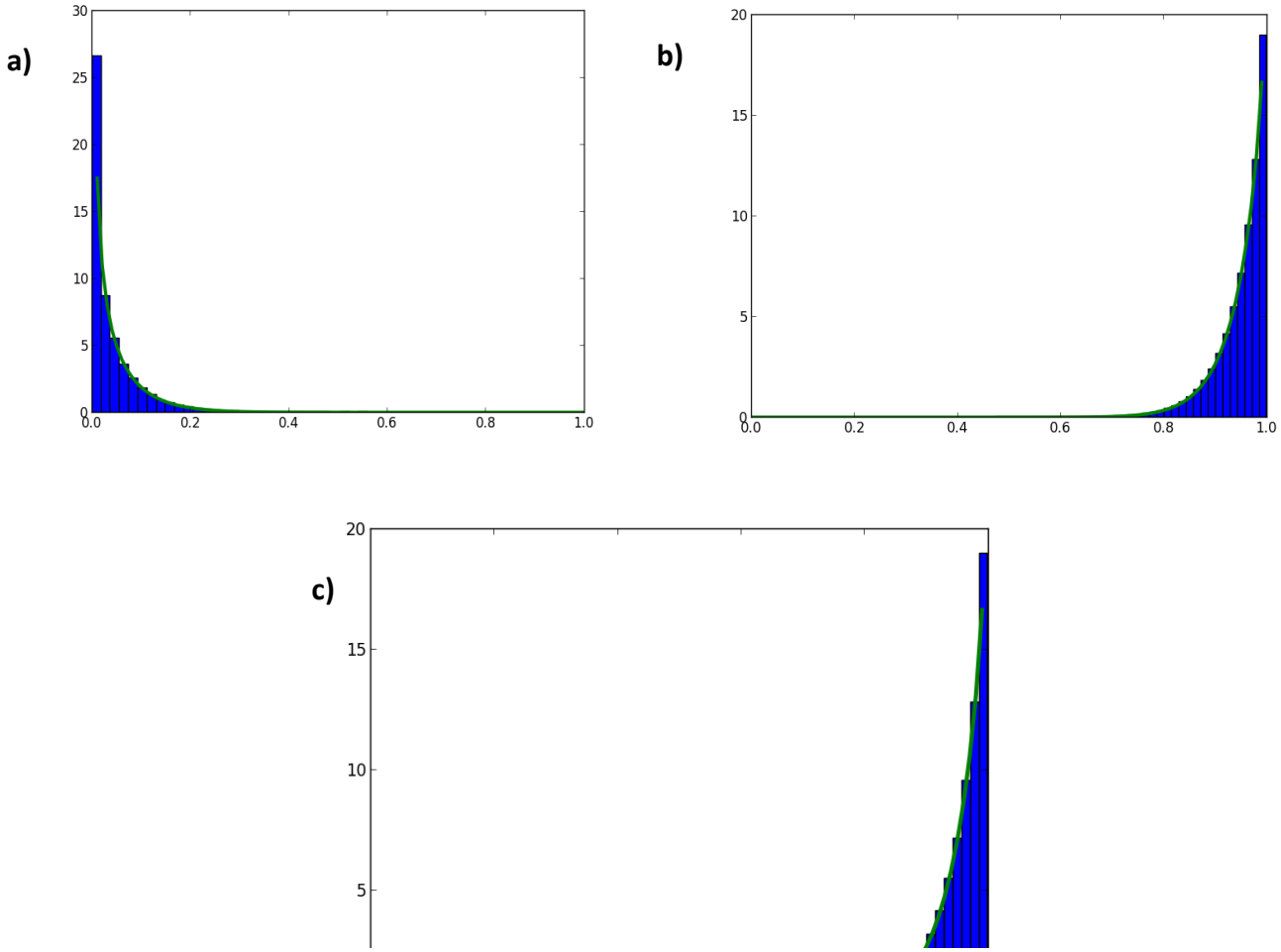

Рис. 23. Гистограммы распределений и плотности для $\boldsymbol{CTR}$ (a), коэффициентов затухания (b) и ставок (c).

Таким образом, алгоритм получения симуляционных данных получен, можно приступать к сравнению двух алгоритмов: базового и учитывающего позиционные эффекты.

3.4.2 Сравнение алгоритма, учитывающего позиционные эффекты и базового алгоритма.

Для тестирования рассматриваемого алгоритма и сравнения его результатов с результатами алгоритма, описанного в п. 2.3 (далее в тексте он также будет упоминаться как "базовый алгоритм"), производилась генерация модельных данных (с разными размерами модельных пулов, при различных параметрах распределений ставок, CTR и т.д.). Затем на этих модельных данных запускались новый и

базовый алгоритмы (при различных ограничениях на денежные средства и покрытие), и сравнивались результаты. Отметим, что схема работы у сравниваемых алгоритмов общая (п. 2.5.1 и п. 3.3.9). Результаты работы обоих алгоритмов имеют общий вид: это списки рекламных объявлений (часть которых могут быть пустыми) с указанием положения этих объявлений в рекламном блоке для каждого запроса из обучающего набора запросов и значения $\lambda_{1\,\text{опт}}$, $\lambda_{2\,\text{опт}}$ и $\lambda_{3\,\text{опт}}$, используемые при работе с новыми запросами. Отличаются максимизируемые критерии (и методы максимизации этих критериев). В рассматриваемом алгоритме это:

$$\overline{CTR}_{pos}(\boldsymbol{T}) = \frac{\sum_{(a,q,p,k)} CTR_{aqpk} \cdot t_{aqpk}}{\sum_{(a,q,p,k)} t_{aqpk}}$$

а в базовом:

$$\overline{CTR}_{no\_pos}(\boldsymbol{T}) = \frac{\sum_{(a,q)} CTR_{aq} \cdot t_{aq}}{\sum_{(a,q)} t_{aq}}$$

Где $CTR_{aq}-$ оценка вероятности клика на -ое рекламное объявление при его показе в рекламном блоке над результатами поиска в ответ на q-ый запрос; $t_{aq}-$ индикатор такого показа. Аналогично, ожидаемое количество денежных средств можно подсчитывать с использованием CTR_{aqpk}:

$$M_{pos}(\boldsymbol{T}) = \sum_{(a,q,p,k)} CTR_{aqpk} \cdot Bid_a \cdot t_{aqpk}$$

либо с использованием непозиционных кликабельностей CTR_{aq}:

$$M_{no\_pos}(\boldsymbol{T}) = \sum_{(a,q)} CTR_{aq} \cdot Bid_a \cdot t_{aq}$$

В силу того, что при генерации модельных данных принималась следующая связь между CTR_{aqpk} и CTR_{aq}: $CTR_{aqpk} = CTR_{aq} \cdot \gamma_p \cdot decay_q^{k-1}$, мы можем легко переходить от позиционных кликабельностей к непозиционным и наоборот при подсчете интегральных величин. Например, по спискам рекламных объявлений, отобранным предыдущим алгоритмом, мы можем посчитать как $\overline{CTR}_{no\_pos}(\boldsymbol{T})$, так и $\overline{CTR}_{pos}(\boldsymbol{T})$.

Схема сравнения текущего и базового такова:

1) Задать параметры модельного набора запросов, сгенерировать его.
2) Задаться ограничениями на ожидаемый суммарный доход и на покрытие.
3) Произвести отбор рекламных объявлений при заданных ограничениях при помощи предыдущего алгоритма; по составленным в результате его работы спискам вычислить среднюю кликабельность $\overline{CTR}_{pos}^0$.
4) Отобрать объявления при заданных ограничениях при помощи нового алгоритма. При этом ожидаемое количество денежных средств должно вычисляться по той же формуле, что и в базовом алгоритме: $M_{no\_pos}(\boldsymbol{T}) = \sum_{(a,q)} CTR_{aq} \cdot Bid_a \cdot t_{aq}$, в противном случае результаты двух алгоритмов будут несравнимы. Действительно, при позиции $k > 1$ $CTR_{aqpk} < CTR_{aq}$, поэтому если вычислять по одному и тому же списку баннеров для спец-размещения величины $M_{pos}(\boldsymbol{T})$ и $M_{no\_pos}(\boldsymbol{T})$, то, скорее всего, окажется $M_{pos}(\boldsymbol{T}) < M_{no\_pos}(\boldsymbol{T})$ (если только этот список не состоит из единственногообъявления: в таком случае $CTR_{aq11} \geq CTR_{aq}$) . По составленным в результате работы

нового алгоритма спискам вычислить среднюю кликабельность $\overline{CTR}_{pos}$.

5) Сравнить $\overline{CTR}^0_{pos}$ и $\overline{CTR}_{pos}$.

Параметры, используемые при генерации некоторых модельных пулов, приведены в Табл.3. (используемые обозначения представлены в разделе 3.4.1):

Номер модельного пула	1	2	3	4	5	6	7
M	400	600	500	500	1400	100	1000
N	120000	180000	230000	230000	500000	240	12400
ads_{min}	5	5	5	5	5	5	5
ads_{max}	40	40	40	40	40	10	40
γ_1	1.065	1.045	1.045	1.045	1.045	1.065	1.065
γ_2	0.92	0.92	0.92	0.95	0.97	0.92	0.92
γ_3	0.83	0.83	0.83	0.902	0.902	0.83	0.83
α_{decay}	12.8	12.8	11.8	14.8	17.8	12.8	12.8
β_{decay}	1.8	1.8	1.1	1.1	0.9	1.8	1.8
μ_{Bid}	1.93	1.93	4.13	4.13	4.13	1.93	1.93
σ_{Bid}	0.47	0.47	0.67	0.67	0.67	0.47	0.47
α_{CTR}	1.7	1.4	1.4	0.5	0.5	1.7	1.7
β_{CTR}	6	9.5	9.5	12.5	12.5	6	6

Табл. 3. Параметры, используемые для генерации модельных симуляционных данных.

Детальные результаты сравнения алгоритмов на этих пулах приведены в Табл.4. Рассмотрим используемые в ней обозначения. δ_M – это относительная погрешность, с которой в обоих алгоритмах выполняется ограничение $M(\boldsymbol{T}) = M_{min}$ при $\lambda_1 > 0$. Обозначения, имеющие верхний индекс "0" относятся к результатам работы базового алгоритма (не учитывающего позиционные эффекты). Характеристика $\overline{CTR}^0_{pos}$ – это среднее значение величин CTR_{aqpk}, полученное в результате отбора базовым алгоритмом объявлений для показа в рекламном блоке над результатами поиска для

запросов из обучающего набора запросов; $\overline{CTR}_{pos}$ – аналогичное среднее значение, но по результатам нового алгоритма. Нас в первую очередь интересует сравнение этих двух величин: оно проведено в последней строчке таблицы. Так как для каждого запроса и объявления известны как значения CTR_{aq}, так и CTR_{aqpk}, то можно также сравнить средние непозиционные кликабельности, получаемые по сравниваемым алгоритмам (это $\overline{CTR}^0_{no\_pos}$ для базового алгоритма и $\overline{CTR}_{no\_pos}$ для нового).

λ_1^0, λ_2^0 и λ_3^0 – оптимальные значения параметров критерия показа, вычисляемые базовым алгоритмом;

λ_1, λ_2 и λ_3 – оптимальные значения параметров критерия показа, вычисляемые новым алгоритмом. Величины $M^0_{no\_pos}$ и M_{no_pos} – ожидаемые суммарные количества денежных средств для базового и нового алгоритма соответственно, вычисляемые в оптимальных точках по кликабельностям CTR_{aq}, не зависящим от позиционных эффектов. $M^0_{no\_pos}$ и M_{no_pos} должны совпадать с M_{min} с относительной погрешностью не более. δ_M

Номер модельного набора	1	2	3	4	5	6	7
покрытие	0.25	0.27	0.33	0.35	0.3	0.35	0.32
M_{min}	1263	1154	18523	7942	19834	281	4326
$\delta_M, \%$	0.8	0.8	0.5	0.2	3	0.5	0.3
λ_1^0	0.04	0.03	0.01	0.02	0.1	0.03	0.04
λ_2^0	0.67	0.43	0.59	0.41	1.6	0.49	0.62
λ_3^0	0.18168	0.0992	0.2587	0.15844	0.716122	0.13954	0.214363
λ_1	0.05	0.04	0.01	0.015	0.1	0.09	0.11
λ_2	0.42	0.3	0.32	0.24	1.2	0.45	0.49
λ_3	0.57237	0.3035	0.62768	0.26439	1.272582	0.47261	0.926016
$M^0_{no\_pos}$	1263	1160	18522	7951	19549	281	4338
M_{no_pos}	1263	1145	18501	7955	19367	279	4320
$\overline{CTR}^0_{no\_pos}$	0.5849	0.4295	0.35556	0.19876	0.180053	0.51489	0.56242

$\overline{CTR}_{no\_pos}$	0.5718	0.4235	0.35326	0.20116	0.17997	0.50305	0.53426
$\overline{CTR}^0_{pos}$	0.4903	0.37324	0.30161	0.18470	0.171925	0.45953	0.4481
$\overline{CTR}_{pos}$	0.5147	0.39194	0.30936	0.18891	0.175112	0.48373	0.46184
$\frac{\overline{CTR}_{pos} - \overline{CTR}^0_{pos}}{\overline{CTR}^0_{pos}}, \%$	10.5	5	2.57	2.28	1.853	5,27	3.1

Табл. 4 Подробные результаты вычислительных экспериментов.

Алгоритм, учитывающий позиционные эффекты, практически всегда оказывается лучше на модельных данных по среднему позиционному CTR, чем алгоритм, не учитывающего позиционные эффекты: в зависимости от характеристик модельных данных улучшение среднего позиционного CTR составляет 1.5 –10% (Рис. 24.).

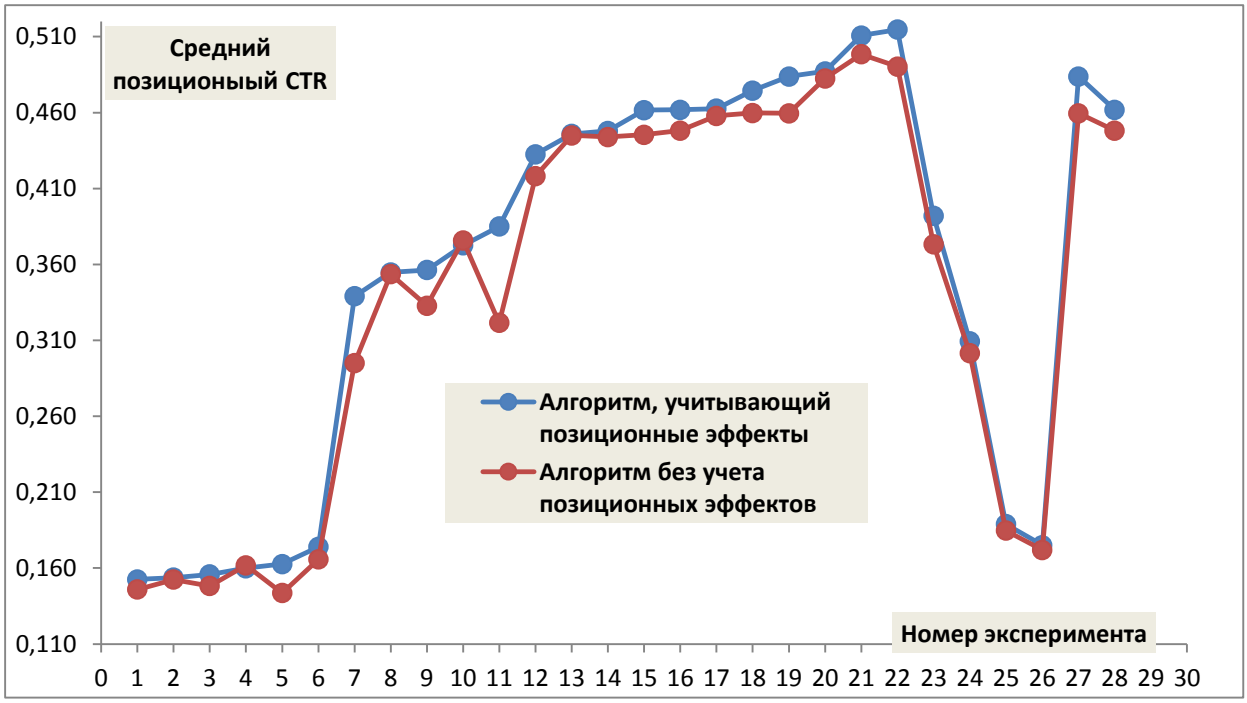

Рис. 24. Сравнение результатов работы базового и позиционного алгоритмов.

ГЛАВА 4. АПРОБАЦИЯ ПОЛУЧЕННОГО ВИДА КРИТЕРИЯ ПОКАЗА В РЕКЛАМНОИ БЛОКЕ НАД РЕЗУЛЬТАТАМИ ПОИСКА. РЕАЛИЗАЦИЯ АЛГОРИТМА ПОДБОРА ПАРАМЕТРОВ КРИТЕРИЯ.

В ходе диссертационного исследования был получен вид функции критерия показа рекламных объявлений над результатами поиска, а так же алгоритм подбора его параметров. Теперь необходимо воспользоваться этим результатом для оптимизации всей системы показов рекламных объявлений. Для разных случаев применения критерия показа в рекламном блоке, а также для выявления значимых результатов апробации, был выработан алгоритм оптимизации системы показов рекламных объявлений в поисковых интернет-системах.

4.1 Алгоритм оптимизации системы показов рекламных объявлений.

Была разработана и реализована схема алгоритма оптимизация системы показов поисковой рекламы (Рис.25.).

Рассмотрим данную схему в общем виде, из каких этапов состоит алгоритм оптимизации показов рекламных объявлений:

1) Отбирается **набор запросов** – информация по историческим запросам пользователей к поисковой интернет-системе «Яндекс». Также на данном этапе происходит отбор объявлений-кандидатов на показ в рекламном блоке [50], [56]. Список поисковых запросов с отобранными кандидатами на показ – это материал для обучения и тестирования новых формул предсказания CTR, а также различных комбинаций показа рекламных объявлений над результатами поиска.

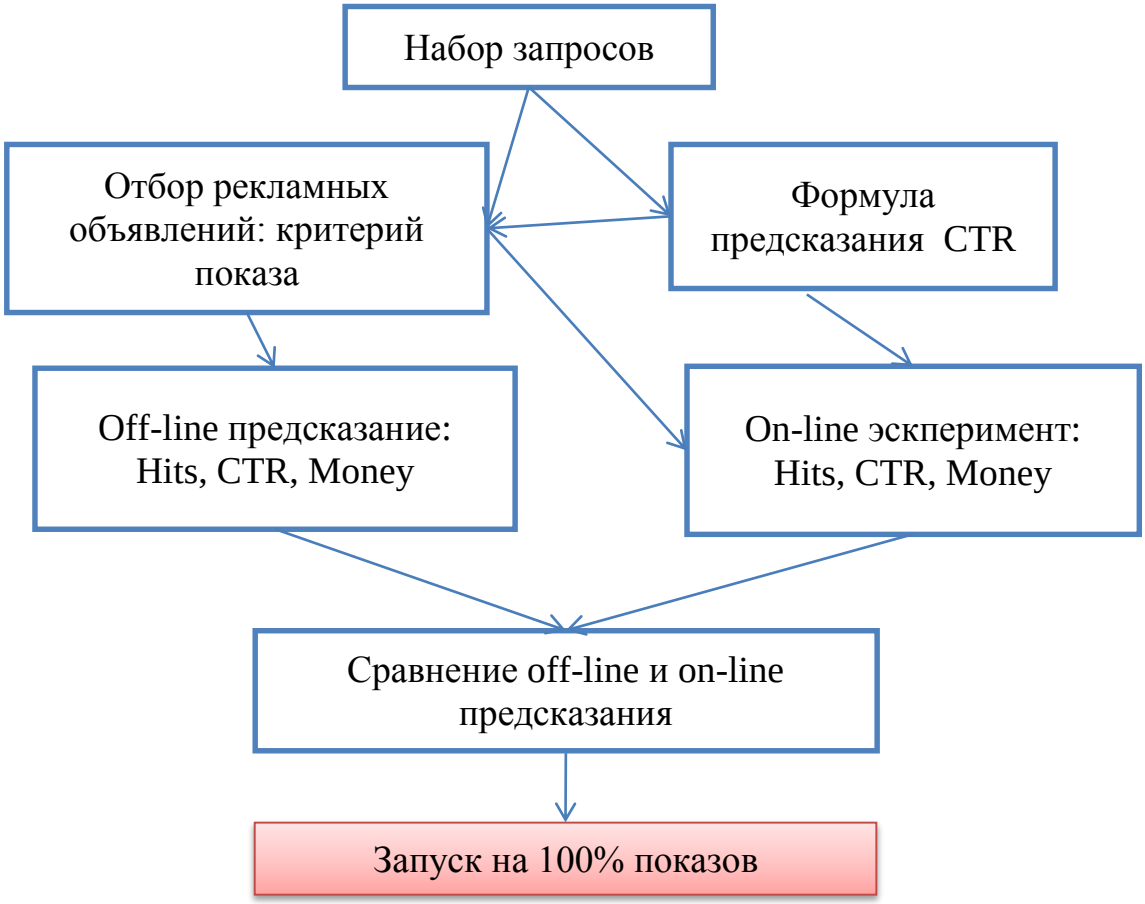

Рис. 25. Общая схема алгоритма оптимизации показов рекламы.

2) Используется **новая формула предсказания CTR** рекламного объявления. Тут возможны разные варианты получения новой формулы, например такие как:

   - добавление новых признаков для использования при уже зафиксированном методе предсказания [17];
   - переобучение метода предсказания на более свежих данных (с тем же набором признаков и методом предсказания) [20];
   - использование нового метода предсказания кикабельности рекламного объявления [8], [34];
   - использование различных кликовых моделей [68].

3) **Отбор рекламных объявлений для показа в рекламном блоке**: для каждого запроса известны объявления-кандидаты на показ

рекламы по данному запросу. Необходимо отобрать те рекламные объявления, которые мы хотим показать над результатами поиска (это можно делать для объявлений как с новой предсказанной кликабельностью, так и с текущим предсказанием CTR). Именно на этом этапе применялись результаты, полученные в ходе диссертационного исследования.

4) Как только мы знаем набор объявлений, которые будут показаны, становится возможным узнать **off-line предсказания** изменения основных показателей (таких как средний CTR, количество запросов с рекламой, суммарный доход от показов и т.д.) для того набора запросов, на котором подбирались пороги по следующим формулам:

$$CTR = \sum_{(a,q)} CTR_{aq}/Events$$

$$M = \sum\nolimits_{(a,q)} Bid_a \cdot CTR_{aq}$$

$Events$ – количество показов рекламных объявлений над результатами поиска.

5) Как только становится возможным для каждого объявления считать новую формулу предсказания CTR, и для неё подобраны новые параметры критерия показа, то можно на части трафика запускать **эксперимент для проверки off-line предсказаний**. Эксперимент проводится на части поискового трафика.

6) Как только эксперимент продлился нужное время для получения значимых изменений в основных характеристиках, то можно **сравнить off-line и on-line показатели** и принять решение о внедрении новой формулы и (или) новых порогов на всём трафике

поисковых запросов. Для определения значимости отклонения основных характеристик используется A/B-testing [24], [44].

Рассмотрим более подробно этап подбора параметров критерия показа объявления в рекламном блоке. Для этой части системы есть более подробная схема (Рис. 26.).

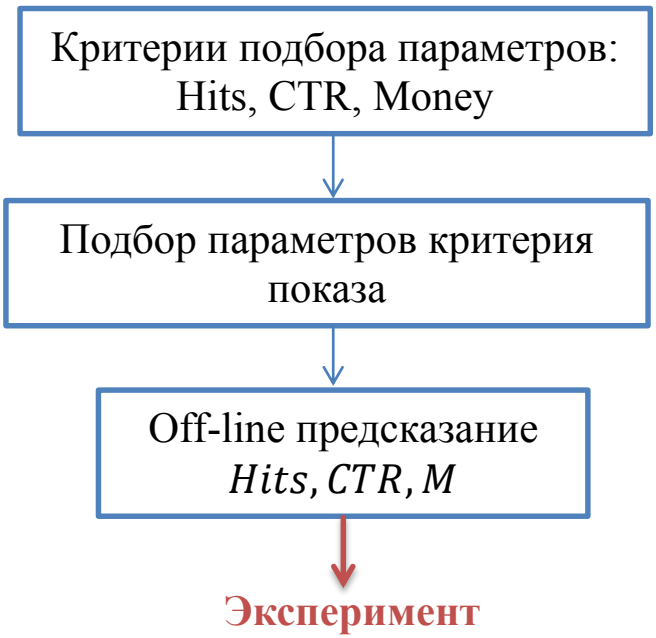

Рис. 26. Схема подбора параметров критерия показа.

Рассмотрим более подробно каждый из пунктов:

- **Критерии подбора параметров критерия показа: $Hits, CTR, Money$.**

В качестве критериев выбраны следующие характеристики:

Hits – ограничение на покрытие рекламы (*см* п. 1.3). Для поисковой системы важно то, насколько много она показывает рекламных объявлений над результатами поиска, так как от этого показателя зависят её основные показатели: во-первых доход (так как рекламный блок над результатами поиска – самое прибыльное место для показа рекламы), а во-вторых – удовлетворение пользователя.

$\boldsymbol{CTR}$ – средняя кликабельность объявлений, показанных над результатами поиска (*см* п. 1.3). При подборе параметров оценивается по более новой формуле предсказания кликабельности конкретных рекламных объявлений. Средняя кликабельность говорит как о качестве самого объявления, так и о том, на сколько реклама релевантна запросам пользователей.

$\boldsymbol{Money}$ – доход поисковой системы – важная составляющая её успешной работы.

До того как будут подобраны параметры критерия показа, необходимо зафиксировать критерии-ограничения по основным характеристикам. В зависимости от этих ограничений можно получить разные постановки задач оптимизации показов рекламы (*см.* п.4.2), например, такие как: $\{Hits = const,\ Money = const,\ CTR \to max\}$, $\{Hits, \% <= -C,\ Money = const,\ CTR \to max\}$. Каждая из постановок задачи оптимизации требует отдельного рассмотрения. На данный момент допустим, что метод подбора параметров критерия показа в рекламном блоке над результатами поиска для каждой из них зафиксирован.

- **Подбор параметров критерия показа, off-line прогноз.**

Как только параметры критерия показа получены и показы рекламы зафиксированы – у нас есть off-line предсказание изменения системы по основным её характеристикам. Также у нас есть возможность запустить on-line эксперимент для проверки полученных off-line предсказаний.

- **Проведение on-line эксперимента.**

На данный момент каждый эксперимент проводится на некоторой доле интернет-пользователей поиска «Яндекса» (обычно эта доля составляет от 1 до 5%): на все их запросы реклама показывается по

новому экспериментальному методу (будь то новая формула предсказания CTR или просто изменение параметров критерия показа в рекламном блоке). В среднем за день это примерно 1.4 миллионов запросов с показом рекламного блока над результатами поиска. Мониторинг эксперимента производится с периодичностью в день (как только накапливаются соответствующие логи показов рекламы), существует две эталонных выборки (с базовой текущей формулой), с которыми сравниваются остальные экспериментальные формулы. Для каждой из характеристик экспериментов считается значимость её отклонения от соответствующей базовой характеристики эталонов. По прошествии (как минимум) недели эксперимента можно говорить о его on-line результатах [24], [44].

- **Сравнение off-line предсказания и результатов on-line экспериментов.**

Надо принимать во внимание то, что у эксперимента на части поискового трафика есть обратная связь от рекламодателей, пользователей и т.д. Из-за этого влияния off-line и on-line результаты могут различаться. Важно определить: значимо они отличаются друг от друга или нет. Если отличие не значимо и (или) результаты эксперимента считаются положительными (достаточными для внедрения на всём трафике поисковых запросов), то текущая формула предсказания кликабельности заменяется на экспериментальную формулу (или происходит изменение параметров критерия показа на новые значения), и весь процесс повторяется снова. Если же on-line предсказания оказываются неудовлетворительными, то новая формула (или новые параметры критерия показа) отвергается и ищется причина несоответствия ожидаемого и действительного.

Одним из наиболее важных и существенных этапов для решения задачи размещения рекламных объявлений на запрос пользователя является подбор параметров показа в рекламном блоке над результатами поиска.

4.2 Различные постановки задачи оптимизации системы показов рекламных объявлений.

Исходя из основных характеристик поисковых интернет-систем, можно выделить несколько типов оптимизационных задач:

$-\{\boldsymbol{Hits = const},\ \boldsymbol{Money = const},\ \boldsymbol{CTR \rightarrow max}\}$

Максимизация средней кликабельности при остальных неизменных параметрах. При данной постановке задачи мы максимизируем эффективность показов рекламы с учётом ограничения по покрытию и по суммарному доходу. Это означает, что мы не хотим показывать над результатами поиска больше рекламы, чем раньше, при этом оставаясь при том же доходе. Однако, среднее качество рекламных объявлений мы хотим улучшить, тем самым количество кликов так же увеличится. Это говорит о том, что пользователи, которым показывается такое же количество рекламы, станут в среднем кликать больше, тем самым повышается их достижение цели по поиску нужной им информации. Для рекламодателей это тоже эффективно: за те же самые денежные средства их бюджетов кликов они получают больше, следовательно средняя стоимость для них рекламной компании уменьшается.

$-\{\boldsymbol{Hits = const},\ \boldsymbol{CTR = const},\ \boldsymbol{Money \rightarrow max}\}$

Максимизация дохода от показов рекламы при ограничении на покрытие и средний CTR. В данной постановке задачи оптимизации мы пытаемся увеличивать доход поисковой системы до тех пор, пока ограничения на покрытие или на средний CTR выполняются. Мы не хотим увеличивать количество поисковых запросов с рекламным

блоком над поисковыми результатами (то есть «зарекрамленность» поисковой выдачи), а также не хотим уменьшать эффективность и привлекательность рекламы для пользователя, фиксируя для этого среднюю кликабельность. Однако при этой постановке задачи страдают рекламодатели: за то же количество кликов они заплатят больше денежных средств (то есть им придётся повысить ставки).

$-\{\boldsymbol{Hits}, \% \leq -\boldsymbol{C},\ \boldsymbol{Money} = \boldsymbol{const},\ \boldsymbol{CTR} \rightarrow \boldsymbol{max}\}$

Максимизация CTR при уменьшении покрытия не больше чем на C процентов относительно текущего состояния, неизменный доход. Мы хотим уменьшить покрытие, при этом эффективность рекламы так же повышается (мы показываем меньше рекламы, при этом показываем только наиболее кликабельную), однако при уменьшении покрытия количество кликов по рекламе может уменьшается, следовательно и суммарный доход может уменьшится. Чтобы этого не происходило ставится ограничение на суммарный доход.

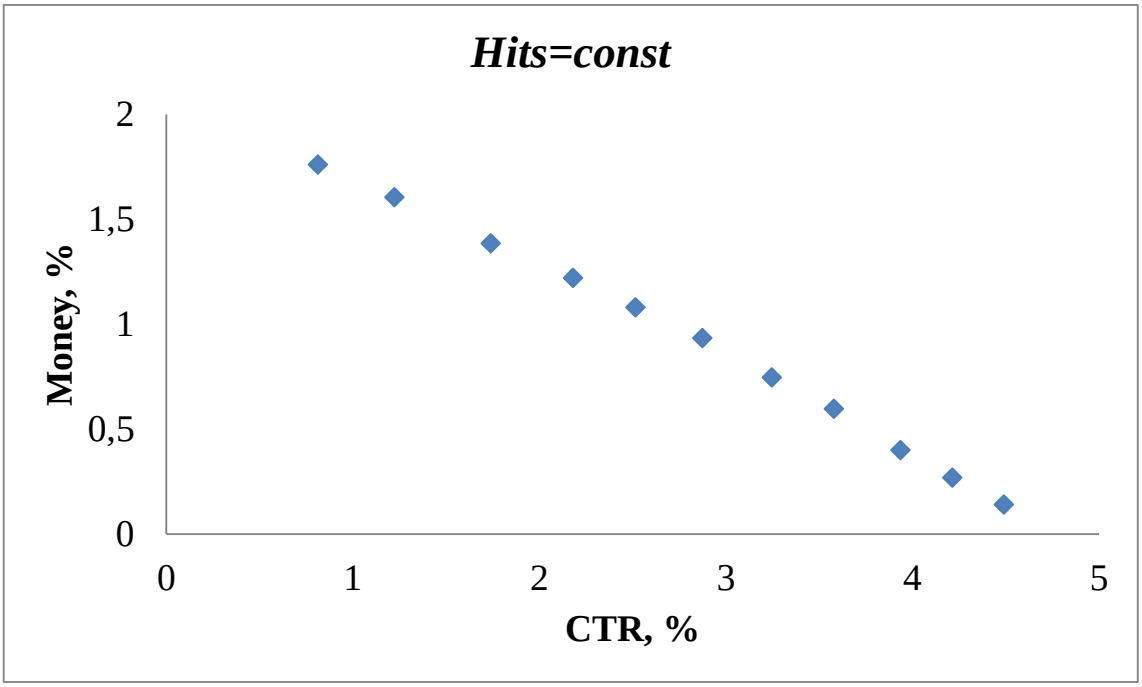

Рис. 27. Изменения средней кликабельности и дохода поисковой системы при разных вариантах $\boldsymbol{\lambda_{1\text{опт}}}$, $\boldsymbol{\lambda_{2\text{опт}}}$ и $\boldsymbol{\lambda_{3\text{опт}}}$

$-\{\boldsymbol{Hits}, \% \leq -\boldsymbol{C},\ \boldsymbol{CTR} = \boldsymbol{const},\ \boldsymbol{Money} \rightarrow \boldsymbol{max}\}$

Максимизация суммарного дохода при увеличении покрытия не больше чем на C процентов относительно текущего состояния, неизменный CTR. Может возникнуть задача увеличения текущего покрытия рекламы, при этом ожидается увеличение количества кликов и, соответственно, дохода. Однако чем больше рекламы мы показываем, тем больше среди показанных рекламных объявлений встречаются объявления с низкой кликабельностью. В связи с этим целесообразно ограничение по среднему CTR рекламных объявлений, показанных в рекламном блоке над результатами поиска.

Пример работы алгоритма для разных постановок задач (Рис. 27.). В данной оптимизационной задаче мы фиксируем покрытие рекламой над результатами поиска и ищем различные варианты изменения средней кликабельности и дохода поисковой системы. На графике (Рис. 27.) каждая точка – это изменение характеристик системы для некоторого набора параметров критерия показа. Видно, что при разных значениях параметров могут решаться разные оптимизационные задачи, например:

- Максимизация дохода при фиксированной средней кликабельности.
- Максимизация средней кликабельности при фиксированном доходе.
- Увеличение дохода на 1%, максимизация средней кликабельности.
- Увеличение средней кликабельности на 3%, максимизация дохода поисковой системы.

Полученный в результате диссертационного исследования вид критерия показа позволяет решать оптимизационные задачи любого из представленных типов.

4.3 Подбор параметров критерия показа.

Как только вид задачи оптимизации определён, то необходимо подобрать оптимальные параметры критерия показа в рекламном блоке над результатами поиска: $\lambda_{1\text{опт}}$, $\lambda_{2\text{опт}}$ и $\lambda_{3\text{опт}}$. Система показов рекламных объявлений в интернет-системе «Яндекс» очень сложная, поэтому подбор параметров on-line на данном этапе мало возможен. В настоящее время алгоритм подбирает $\lambda_{1\text{опт}}$, $\lambda_{2\text{опт}}$ и $\lambda_{3\text{опт}}$ на наборе запросов, состоящим из 100 000 случайных запросов из поисковых логов за неделю данных. Из этих 100 000 запросов только на части будет показана реклама над результатами поиска (ограничение на покрытие), однако чтобы получить значения $\lambda_{1\text{опт}}$, $\lambda_{2\text{опт}}$ и $\lambda_{3\text{опт}}$ нужно просмотреть весь набор запросов и соответствующих им объявлений-кандидатов на показ.

Для подбора значений параметров $\lambda_{1\text{опт}}$, $\lambda_{2\text{опт}}$ и $\lambda_{3\text{опт}}$ был ***реализован комплекс программ***, включающий в себя следующие компоненты:

1) **QueryPool.py** – выбор из лога показов рекламных объявлений случайного ***набора запросов*** необходимого размера. На вход программе даются логи рекламных показов: временная последовательность запросов пользователей, на которые было показано хотя бы одно рекламное объявление. Пользовательские запросы нужны для того, чтобы собрать ***обучающий материал*** для подбора параметров. За день в поисковую систему поступают миллионы запросов, это количество слишком велико для обучения, поэтому возникает необходимость выбрать 10^5 случайных из них за неделю данных, что и делает эта программа.
2) **GetCandidatesForQuery.py** – для каждого из запросов из базы данных рекламных объявлений выбираются ***соответствующие ему рекламные объявления***. Для каждого из запросов, которые

были собраны программой **QueryPool.py**, производятся следующие действия:

- Из запроса выделяются ключевые слова – слова, несущие основную смысловую нагрузку.
- Из ключевых слов (если их несколько) составляются ключевые фразы.
- По ключевым фразам из всего набора рекламных объявлений выбираются именно те, которые торгуются по соответствующим фразам из запроса. Таким образом, для каждого запроса отбираются соответствующие ему кандидаты на показ в рекламном блоке над результатами поиска.
- Объявления-кандидаты некоторым образом фильтруются (чтобы не было повторяющихся объявлений, объявлений от одного рекламодателя и т.д.)
- Для каждого объявления-кандидата известна его ставка Bid и вычисляется предсказание вероятности клика CTR.

В результате работы программы получается ***обучающий материал для подбора параметров критерия показа.***

3) **OptimalThresholdParameters.py** – реализация ***алгоритма подбора значений параметров*** критерия показа в зависимости от поставленной оптимизационной задачи. На вход программа получает:

- Набор запросов с объявлениями-кандидатами, для каждого из которых известны значения Bid и CTR, которые были получены предыдущей программой **GetCandidatesForQuery.py**
- Критерий, который необходимо максимизировать: $CTR, Money$.
- Ограничения поисковой системы: $CTR, Money, Hits, k, Events$.

После того, как задана конкретная задача оптимизации с ограничениями, выполняется поиск соответствующих значений $\lambda_{1\text{опт}}$, $\lambda_{2\text{опт}}$ и $\lambda_{3\text{опт}}$ (Рис.1.). Полный цикл подбора порогов для набора запросов занимает от 6 до 9 часов (в зависимости от шага перебора λ_1 и λ_2).

Реализация комплекса программ была выполнена на языке программирования Python. Из-за больших объёмов данных (логов показов рекламы, обучающего набора запросов и объявлений-кандидатов на показ) возникла необходимость использования ***распределённых вычислений*** [60] на MapReduce [26]: вычисления оптимальных параметров происходят на порядок быстрее.

Важно заметить, что данный комплекс программ позволяет подбирать значения параметров критерия показа для любой части поискового трафика. Если возникнет задача подбора значений параметров критерия показа отдельно для регионов или для дней недели, или ещё по каким-либо разделениям (по пользователям, рекламодателям, гео-таргетинг, временной таргетинг и т.п.), то это не представляется проблемой. Таким образом, программный комплекс является масштабируемым.

4.4 Проведение on-line эксперимента, внедрение на 100% поискового трафика.

После того как значения параметров критерия показа $\lambda_{1\text{опт}}$, $\lambda_{2\text{опт}}$ и $\lambda_{3\text{опт}}$ подобраны, необходимо запустить эксперимент на реальных пользователях. Как было сказано в п. 4.1, эксперимент обычно проводится на 1-5% поискового трафика. В данном случае решалась задача максимизация средней кликабельности при ограничении на доход поисковой системы и доли запросов с рекламой над результатами поиска.

В среднем на каждый запрос пользователя отбирается 50 объявлений-кандидатов, для каждого из которых нужно вычислить критерий показа в рекламном блоке над результатами поиска:

$$F_{aq} = CTR_{aq} + \lambda_{1\text{опт}} \cdot Bid_a \cdot CTR_{aq} - \lambda_{2\text{опт}}$$

Кликабельность объявления CTR_{aq} и его ставка Bid_a уже известны, таким образом, чтобы посчитать критерий показа, нужно произвести всего лишь ряд элементарных математических операций. После того как полный рекламный блок для показа над результатами поиска сформирован, для него считается суммарный критерий: $R_q = \sum F_{aq}$, который сравнивается с параметром $\lambda_{3\text{опт}}$: если $R_q < \lambda_{3\text{опт}}$, то рекламные объявления над результатами поиска не показываются, иначе – показывается весь рекламный блок.

Для одного запроса отбор объявлений для показа выполняется за сотые доли секунды: это соответствует требованиям производительности поисковой системы (необходим быстрый ответ на запрос пользователя).

On-line эксперимент проводился 10 дней на 2% поискового трафика, получены результаты изменения средней кликабельности по рекламному блоку над результатами поиска относительно эталонного эксперимента (Табл.5.).

После того, как на on-line эксперименте был получен средний прирост CTR 8%, было решено внедрить данный вид критерия показа в рекламном блоке над результатами поиска для всей системы показов рекламы компании «Яндекс». После внедрения некоторое время проводился «инверсный» эксперимент со старой формулой. Результаты сравнения отбора с использованием нового вида критерия показа и «инверсного» эксперимента по средней кликабельности в

процентах можно увидеть в Табл.5. Средний прирост кликабельности составил 8.2%.

№ дня	1	2	3	4	5	6	7	8	9	10	Среднее
эксперимент	7.9%	8.1%	8.2%	7.8%	8.3%	7.7%	7.6%	7.8%	8.5%	8.2%	**8%**
внедрение	7.9%	8.8%	7.7%	8.1%	8.7%	7.5%	7.5%	8.1%	8.4%	8.3%	**8.1%**

Табл.5. Дневная динамика изменения средней кликабельности по сравнению с эталонным и инверсным экспериментами.

Данный вид критерия показа в рекламном блоке над результатами поиска используется на данный момент в системе показов рекламных объявлений компании «Яндекс».

ЗАКЛЮЧЕНИЕ.

Основные результаты, полученные лично соискателем, и их научная новизна заключаются в том, что:

1. На основе проведённого критического анализа существующих подходов и методов разработана **новая модель** показов рекламных объявлений в поисковых системах, а также получено её **математическое описание**.
2. В ходе решения задачи оптимизации рекламных показов выявлен **новый вид критерия**, с помощью которого производится отбор кандидатов на показ в рекламном блоке над результатами поиска, включающий в себя эффективность показа объявления и его доходность для поисковой системы.
3. Получен алгоритм **подбора параметров критерия** показа рекламных объявлений. С помощью этих параметров стало возможным работать с новыми поисковыми запросами. Получена усовершенствованная модификация базового алгоритма подбора параметров критерия показа рекламных объявлений. Модифицированный алгоритм изменён для учёта позиционных эффектов в показе рекламного блока над результатами поиска.
4. Написан **комплекс программ**, основной частью которого является **алгоритм подбора параметров** критерия показа. Алгоритм показал высокую эффективность, масштабируемость и быстродействие. По результатам тестирования нового вида критерия показа на on-line эксперименте было решено **использовать предложенный вид критерия** для всего потока запросов поисковой системы «Яндекс».

СПИСОК ЛИТЕРАТУРЫ.

1. Бауман К.Е., Топинский В.А., Корнетова А.Н., Хакимова Д.А., Оптимизация прогноза вероятности клика по контекстной рекламе на поисковой системе «Яндекса» // Научно-Техническая Информация. Серия 2. Информационные процессы и системы, 2013.– №. 4.– С. 1-8.
2. Корнетова А. Н., Червоненкис А. Я. Оптимизация показов рекламы в поисковых системах //Проблемы управления. – 2013. – №. 1. – С. 40-49.
3. Кун Г. У., Таккер А. У. Линейные неравенства и смежные вопросы.- М.: Изд-во иностр. лит. – 1959.
4. Поляк Б. Т. Введение в оптимизацию. – М.: Наука. Гл. ред. физ.-мат. лит. - 1983.
5. Сорокина А. Н. Алгоритм размещения рекламных объявлений над результатами поиска, максимизирующий доход поисковой системы //Информационные процессы. – 2014. – Т. 14. – №. 1. – С. 108-116.
6. Agarwal A., Hosanagar K., Smith M. D. Location, location, location: An analysis of profitability of position in online advertising markets //Journal of marketing research. – 2011. – V. 48. – №. 6. – P. 1057-1073.
7. Agarwal D. K., Jung D. M., Li S. M., Mahdian M., McAfee R. P., Ravikumar S., Reiley D. System and method for exploring new sponsored search listings of uncertain quality. Patent Application 12/700,530 USA. – 2010.
8. Agarwal D. Prediction of click through rates using hybrid kalman filter-tree structured markov model classifiers. Patent Application 7680746 USA. – 2010.
9. Ashkan A., Clarke C. L. A. Impact of query intent and search context on clickthrough behavior in sponsored search //Knowledge and information systems. – 2013. – V. 34. – №. 2. – P. 425-452.

10. Attenberg J., Pandey S., Suel T. Modeling and predicting user behavior in sponsored search //Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2009. – P. 1067-1076.
11. Battelle J. The Search: How Google and Its Rivals Rewrote the Rules of Business andTransformed Our Culture. – Penguin, 2005.
12. Bauman K. E., Kornetova A. N., Topinskii V. A., Khakimova D. A. Optimization of click-through rate prediction in the Yandex search engine //Automatic Documentation and Mathematical Linguistics. – 2013. – V. 47. – №. 2. – P. 52-58.
13. Briggs R., Hollis N. Advertising on the web: is there response before click-through? //Journal of Advertising Research. – 1997. – V. 37. – P. 33-46.
14. Broder A. Z., Ciccolo P., Fontoura M., Gabrilovich E., Josifovski V., Riedel L. Search advertising using web relevance feedback //Proceedings of the 17th ACM conference on Information and knowledge management. – ACM, 2008. – P. 1013-1022.
15. Brooks N., Magun H. Navigational behaviour and sponsored search advertising //International Journal of Electronic Business. – 2008. – V. 6. – №. 2. – P. 132-148.
16. Chakrabarti D., Agarwal D., Josifovski V. Contextual advertising by combining relevance with click feedback //Proceedings of the 17th international conference on World Wide Web. – ACM, 2008. – P. 417-426.
17. Chen Y., Kapralov M., Pavlov D., Canny J. Factor Modeling for Advertisement Targeting //NIPS. – 2009. – V. 9. – P. 324-332.
18. Cheng H., Cantu-Paz E. Personalized click prediction in sponsored search //Proceedings of the third ACM international conference on Web search and data mining. – ACM, 2010. – P. 351-360.

19. Chervonenkis A., Sorokina A., Topinsky V. A. Optimization of ads allocation in sponsored search //Proceedings of the 22nd international conference on World Wide Web companion. – International World Wide Web Conferences Steering Committee, 2013. – P. 121-122.
20. Chu W., Zinkevich M., Li L., Thomas A., Tseng, B. Unbiased online active learning in data streams //Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2011. – P. 195-203.
21. Ciaramita M., Murdock V., Plachouras V. Online learning from click data for sponsored search //Proceedings of the 17th international conference on World Wide Web. – ACM, 2008. – P. 227-236.
22. Clark D. Start-up plans Internet search service tying results to advertising spending //The Wall Street Journal. – 1998.
23. Craswell N., Zoeter O., Taylor M., Ramsey B. An experimental comparison of click position-bias models //Proceedings of the 2008 International Conference on Web Search and Data Mining. – ACM, 2008. – P. 87-94.
24. Crook T., Frasca B., Kohavi R., Longbotham R. Seven pitfalls to avoid when running controlled experiments on the web //Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2009. – P. 1105-1114.
25. De Filippi G. Keywords auto-segmentation and auto-allocation system to increase search engines income. Patent Application 11/382,276 USA. – 2006.
26. Dean J., Ghemawat S. MapReduce: simplified data processing on large clusters //Communications of the ACM. – 2008. – V. 51. – №. 1. – P. 107-113.
27. Dembczynski K., Kotlowski W., Weiss D. Predicting ads' click-through rate with decision rules ranking //Online Advertising. – 2008.

28. Dudley B. Microsoft Touts Ad-Selling System as Step Ahead of its Competitors //Seattle Times. – 2005.
29. Edelman B., Ostrovsky M., Schwarz M. Internet advertising and the generalized second price auction: Selling billions of dollars worth of keywords. – National Bureau of Economic Research, 2005. – №. w11765.
30. Fain D. C., Pedersen J. O. Sponsored search: A brief history //Bulletin of the American Society for Information Science and Technology. – 2006. – V. 32. – №. 2. – P. 12-13.
31. Feng J., Bhargava H. K., Pennock D. Comparison of allocation rules for paid placement advertising in search engines //Proceedings of the 5th international conference on Electronic commerce. – ACM, 2003. – P. 294-299.
32. Feng J., Bhargava H. K., Pennock D. M. Implementing sponsored search in web search engines: Computational evaluation of alternative mechanisms //INFORMS Journal on Computing. – 2007. – V. 19. – №. 1. – P. 137-148.
33. Ghose A., Yang S. An empirical analysis of search engine advertising: Sponsored search in electronic markets //Management Science. – 2009. – V. 55. – №. 10. – P. 1605-1622.
34. Graepel T., Candela J. Q., Borchert T., Herbrich R. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine //Proceedings of the 27th International Conference on Machine Learning (ICML-10). – 2010. – P. 13-20.
35. Granka L. A., Joachims T., Gay G. Eye-tracking analysis of user behavior in WWW search //Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. – ACM, 2004. – P. 478-479.

36. Green D. The evolution of Web searching //Online Information Review. – 2000. – V. 24. – №. 2. – P. 124-137.
37. Gupta S., Bilenko M., Richardson M. Catching the drift: learning broad matches from clickthrough data //Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2009. – P. 1165-1174.
38. Herbrich R., Graepel T., Obermayer K. Large margin rank boundaries for ordinal regression //Advances in Neural Information Processing Systems. – 1999. – P. 115-132.
39. Hillard D., Schroedl S., Manavoglu E., Raghavan H., Leggetter, C. Improving ad relevance in sponsored search //Proceedings of the third ACM international conference on Web search and data mining. – ACM, 2010. – P. 361-370.
40. Jansen B. J., Mullen T. Sponsored search: an overview of the concept, history, and technology //International Journal of Electronic Business. – 2008. – V. 6. – №. 2. – P. 114-131.
41. Jansen B. J., Sobel K., Zhang M. The brand effect of key phrases and advertisements in sponsored search //International Journal of Electronic Commerce. – 2011. – V. 16. – №. 1. – P. 77-106.
42. Joachims T., Granka L., Pan B., Hembrooke H., Gay G. Accurately interpreting clickthrough data as implicit feedback //Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. – ACM, 2005. – P. 154-161.
43. Joachims T. Optimizing search engines using clickthrough data //Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2002. – P. 133-142.
44. Kohavi R., Crook T., Longbotham R., Frasca B., Henne R., Ferres J. L., Melamed T. Online experimentation at Microsoft //Data Mining Case Studies. – 2009. – P. 11.

45. Kuhn H. W. The Hungarian method for the assignment problem //Naval research logistics quarterly. – 1955. – V. 2. – №. 1. – P. 83-97.
46. Lacerda A., Cristo M., Gonçalves M. A., Fan W., Ziviani N., Ribeiro-Neto B. Learning to advertise //Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. – ACM, 2006. – P. 549-556.
47. Lahaie S., Pennock D. M. Revenue analysis of a family of ranking rules for keyword auctions //Proceedings of the 8th ACM conference on Electronic commerce. – ACM, 2007. – P. 50-56.
48. Lee K., Seda C. Search engine advertising: buying your way to the top to increase sales. – New Riders, 2009.
49. Liu T. Y., Xu J., Qin T., Xiong W., Li H. Letor: Benchmark dataset for research on learning to rank for information retrieval //Proceedings of SIGIR 2007 workshop on learning to rank for information retrieval. – 2007. – P. 3-10.
50. Mehta A., Saberi A., Vazirani U., Vazirani V. Adwords and generalized online matching //Journal of the ACM (JACM). – 2007. – V. 54. – №. 5. – P. 22.
51. Pin F., Key P. Stochastic variability in sponsored search auctions: observations and models //Proceedings of the 12th ACM conference on Electronic commerce. – ACM, 2011. – P. 61-70.
52. Radlinski F., Broder A., Ciccolo P., Gabrilovich E., Josifovski V., Riedel L. Optimizing relevance and revenue in ad search: a query substitution approach //Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. – ACM, 2008. – P. 403-410.
53. Regelson M., Fain D. Predicting click-through rate using keyword clusters //Proceedings of the Second Workshop on Sponsored Search Auctions. – 2006. – V. 9623.

54. Reiley D. H., Li S. M., Lewis R. A. Northern exposure: A field experiment measuring externalities between search advertisements //Proceedings of the 11th ACM conference on Electronic commerce. – ACM, 2010. – P. 297-304.
55. Richardson M., Dominowska E., Ragno R. Predicting clicks: estimating the click-through rate for new ads //Proceedings of the 16th international conference on World Wide Web. – ACM, 2007. – P. 521-530.
56. Rusmevichientong P., Williamson D. P. An adaptive algorithm for selecting profitable keywords for search-based advertising services //Proceedings of the 7th ACM Conference on Electronic Commerce. – ACM, 2006. – P. 260-269.
57. Schroedl S., Kesari A., Neumeyer L. Personalized ad placement in web search //Proceedings of the 4th Annual International Workshop on Data Mining and Audience Intelligence for Online Advertising (AdKDD), Washington USA. – 2010.
58. Sheth A., Avant D., Bertram C. System and method for creating a semantic web and its applications in browsing, searching, profiling, personalization and advertising. Patent № 6311194 USA. – 2001.
59. Snyder B. Doubleclick revamps in growth spurt //Advertising Age. – 1997. – V. 68. – №. 41. – P. 38.
60. Tanenbaum A. S., Maarten “van” Steen. Distributed systems. – Upper Saddle River : Prentice Hall, 2002.
61. Trofimov I., Kornetova A., Topinskiy V. Using boosted trees for click-through rate prediction for sponsored search //Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy. – ACM, 2012. – P. 2.
62. Wang F., Zhang X. P. S., Ouyang M. Does advertising create sustained firm value? The capitalization of brand intangible //Journal of the Academy of Marketing Science. – 2009. – V. 37. – №. 2. – P. 130-143.

63. Wang X., Li W., Cui Y., Zhang R. B., Mao J. Click-through rate estimation for rare events in online advertising //Online Multimedia Advertising: Techniques and Technologies. – 2011. – P. 1.
64. Xin X., King I., Agrawal R., Lyu M. R., Huang H. Do ads compete or collaborate: designing click models with full relationship incorporated //Proceedings of the 21st ACM international conference on Information and knowledge management. – ACM, 2012. – P. 1839-1843.
65. Yan J., Liu N., Wang G., Zhang W., Jiang Y., Chen Z. How much can behavioral targeting help online advertising? //Proceedings of the 18th international conference on World wide web. – ACM, 2009. – P. 261-270.
66. Zhang W. V., Jones R. Comparing click logs and editorial labels for training query rewriting //WWW 2007 Workshop on Query Log Analysis: Social And Technological Challenges. – 2007.
67. Zhu Y., Wang G., Yang J., Wang D., Yan J., Hu J., Chen Z. Optimizing search engine revenue in sponsored search //Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. – ACM, 2009. – P. 588-595.
68. Zhu Z. A., Chen W., Minka T., Zhu C., Chen Z. A novel click model and its applications to online advertising //Proceedings of the third ACM international conference on Web search and data mining. – ACM, 2010. – P. 321-330.
69. Контекстная реклама в Интернете [Электронный ресурс]. /Режим доступа. 2009. URL: http://reklama_proektov/kontekstnaya-reklama-internete.html (дата обращения: 02. 2013).
70. Развитие интернета в регионах России [Электронный ресурс] // Яндекс сегодня. 2013. URL: http://company.yandex.ru/researches/reports/2013/ya_internet_regions_2013.xml (дата обращения: 05. 2013).

71. Newcomb Ask Jeeves Bows New Sponsored Listings [Электронный ресурс] / ClickZ Internet Advertising News. 2005. URL: http://www.clickz.com/news/article.php/3524151 (дата обращения: 02. 2012).
72. Engine Sells Results, Draws Fire [Электронный ресурс] / News.com. 1996. URL: http://www.news.com/News/Item/Textonly/0 (дата обращения: 02. 2012).
73. Another Engine Takes Ads by the Click [Электронный ресурс] / News.com. 1996. URL: http://www.news.com/News/Item/0 (дата обращения: 03. 2012).
74. Direct Response Marketing [Электронный ресурс] / Wikipedia. 2012. URL:http://en.wikipedia.org/wiki/Direct_response_marketing (дата обращения: 05. 2013).

ПРИЛОЖЕНИЕ 1. АКТЫ О ВНЕДРЕНИИ.

Исх. _____
от «2» марта 2014 г.

СПРАВКА

о внедрении результатов диссертационной работы

Сорокиной А.Н.

«Оптимизация показов рекламных объявлений в поисковых интернет-системах: разработка методологии подбора порогов входа в рекламный показ»

по специальности 05.13.18. – Математическое моделирование, численные методы и комплексы программ на соискание учёной степени кандидата технических наук

Основные результаты и материалы диссертационного исследования Сорокиной А.Н. внедрены в учебный процесс Национального исследовательского университета Высшая школа экономики по подготовке студентов высшего образования в области анализа интернет-данных по направлению 010400.68 Прикладная математика и информатика в рамках курсов лекций и практических занятий по дисциплинам «Современные методы анализа данных» и «Научный семинар «Анализ интернет-данных».

Внедрение результатов диссертационного исследования в учебный процесс позволило повысить теоретический и практический уровень знаний студентов в области методов анализа интернет-данных на примере изучения механизмов показа рекламы в поисковых интернет-системах.

Зав. отделения прикладной математики и информатики, д. ф.-м.н., профессор — Кузнецов С.О.

Зав. базовой кафедрой Яндекс, д. ф.-м.н., профессор — Аржанцев И.В.

Яндекс

ООО «ЯНДЕКС»
ул. Льва Толстого, 16
Москва, Россия, 119021
Тел.: +7 (495) 739-70-00
Факс: +7 (495) 739-70-70
yandex.ru
info@yandex-team.ru

Исх. _____

от «8» мая 2014 г.

СПРАВКА

о внедрении результатов кандидатской диссертационной работы Сорокиной А.Н.

«Оптимизация показов рекламных объявлений в поисковых интернет-системах: разработка методологии подбора порогов входа в рекламный показ».

Настоящая справка выдана Сорокиной Анне Николаевне в том, что основные результаты её диссертационной работы на соискание учёной степени кандидата технических наук внедрены в ООО «Яндекс».

Основной задачей, которой занималась Сорокина А.Н., была оптимизация показов рекламных объявлений над результатами поисковых результатов для запросов пользователей.

В систему показов рекламных объявлений ООО «Яндекс» были внедрены следующие основные результаты, напрямую связанные с темой кандидатской диссертации Сорокиной А.Н.:

- Определение основных показателей эффективности показов рекламных объявлений над результатами поиска, выявление ограничений, которым должна удовлетворять система рекламных показов.
- Решение основных задач по оптимизации показов рекламных объявлений происходит с использованием предложенного вида функции порога входа в рекламный показ.
- Реализован алгоритм подбора параметров пороговой функции в зависимости от поставленной оптимизационной задачи, что позволило повысить производительность и эффективность системы.

Сорокиной А.Н. была проведена апробация предложенных методик и алгоритмов на реальных запросах пользователей. Эксперимент показал преимущество разработанного инструментария: простой вид функции порога повысил интерпретируемость полученных результатов, а также позволил устанавливать разные пороговые значения для отдельных частей потока запросов, что удобно для управления качеством рекламной выдачи.

Моделирование показов рекламных объявлений, предложенное Сорокиной А.Н. в диссертационном исследовании, активно используется в исследовательской и практической деятельности компании. Внедрение результатов диссертационного исследования позволило повысить эффективность показов рекламных объявлений над результатами поиска на 8% и продолжает использоваться для дальнейшей оптимизации системы по различным критериям.

Начальник исследовательской службы в монетизации Топинский В.А.