Московский Авиационный Институт

(Научный Исследовательский Институт)

Факультет прикладной математики и физики

Кафедра вычислительной математики и программирования

Курсовая работа по курсу «Информационный поиск» по теме: «Машинное обучение для задачи ранжирования»

Выполнил: Ефименко Н. А.

Группа: М8О-206М

Преподаватель: Калинин А. Л.

Москва, 2020 г.

ВВЕДЕНИЕ

В современном мире важно развивать технологии релевантного информационного поиска. Это помогает делать жизнь проще, облегчая поиск необходимой информации. В качество поиска входит множество элементов, в то числе и ранжирование. Благодаря использованию машинного обучения в решении данной задачи, значительно повышается автоматизация процесса и, как следствие скорость её выполнения, а также адаптивная оценка в зависимости от условий.

Рассмотрим в чём заключается задача ранжирования, какие существуют традиционные алгоритмы подхода к проблеме. В чём заключается машинное обучение, как оно помогает при ранжировании.

ОСНОВНАЯ ЧАСТЬ

1 Ранжирование

1.1 Постановка задачи

Для начала рассмотрим, что такое ранжирование и какую задачу оно решает. После принятия поисковой системой запроса пользователя она должна найти все подходящие страницы и упорядочить их по принципу наибольшего соответствия запросу. Ранжирование — сортировка сайтов в поисковой выдаче, применяемая в поисковых системах.

Чтобы оценить качество ранжирования, необходимо иметь некоторый «эталон», с которым можно было бы сравнить результаты алгоритма. Его можно получить двумя способами: на основе статистичеких данных или на основе экспертной оценки.

1.2 Стандартные методы ранжирования

Найденные документы ранжируются по набору формальных признаков, что позволяет получить статистически приемлемые результаты. К ним (признакам) относятся:

- релевантность;
- авторитетность;
- актуальность;
- цитируемость;
- поведенческие факторы.

Кратко рассмотрим такую схему ранжирования, как TF-IDF. Это статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Этот метод заключается в том, что каждому термину, встречающемуся в документе, присваивается вес, зависящий от количества его появлений в документе:

$$tf(t,d) = \frac{n_t}{m},$$

где n_t — число вхождений термина t в документ, m — общее количество слов.

Затем для коррекции веса терма используется документная частота. Обратная документная частота имеет вид:

$$idf(t, D) = log\frac{D}{df}, \text{где} D - \text{общееколичестводокументов.}$$

После этого комбинируются частота термина в документе (tf) и обратная документная частота (idf) для получения веса каждого термина в каждом документе по формуле:

$$tf\text{-}idf(t, d, D) = tf(t, d) \times idf(t, D)$$

Таким образом, релевантность документа d равна сумме вхождений всех терминов запроса в этот документ:

$$Score(q, d) = \sum\big(tf\text{-}idf(t, d, D)\big)$$

2 Машинное обучение

2.1 Виды

Машинное обучение – класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач. Машинное обучение делится на четыре основных вида:

- Классическое обучение:
  - с учителем;
  - без учителя.
- Обучение с подкреплением.
- Ансамблевые методы
- Нейросети и глубокое обучение.

Разберём подвиды обучения с учителем, так как они применяются при обучении ранжированию.

Обучение с учителем – один из способов машинного обучения, в ходе которого испытуемая система принудительно обучается с помощью примеров «стимул-реакция».

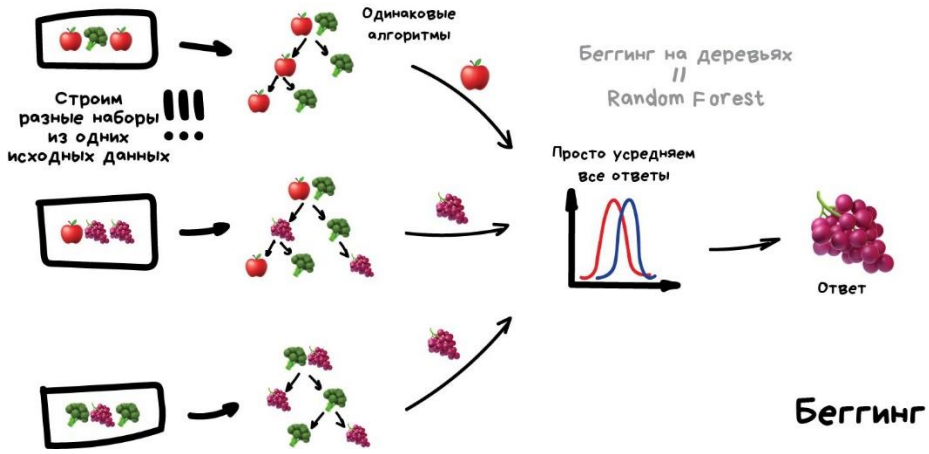
В обучение без учителя испытуемая система, наоборот, обучается выполнять поставленную задачу без вмешательства со стороны экспериментатора. Как правило, это пригодно только для задач, в которых известны описания множества объектов обучающей выборки, и требуется обнаружить внутренние взаимосвязи, зависимости, закономерности, существующие между объектами.

С учителем обучение происходит быстрее и точнее, потому его используют намного чаще. Эти задачи делятся на два типа: классификация — предсказание категории объекта, и регрессия — предсказание места на числовой прямой.

Классификация разделяет объекты по заранее известному признаку. К алгоритмам классификации относятся: деревья решений, SVM и другие. Основная идея метода — перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. В дереве решений идея другая – автоматическое разделение всех данных по вопросам, ответы на которые «да» или «нет». В чистом виде деревья используют редко, но их ансамбли лежат в основе крупных систем и зачастую работают лучше нейросетей. Например, в компании Яндекс, именно они выполняют ранжирование. Ансамбли – это объединение некоторого количества алгоритмов, которые учатся исправлять ошибки друг друга, что повышает качество работы. Разделяют три способа конструирования ансамблей: стекинг, беггинг и бустинг.

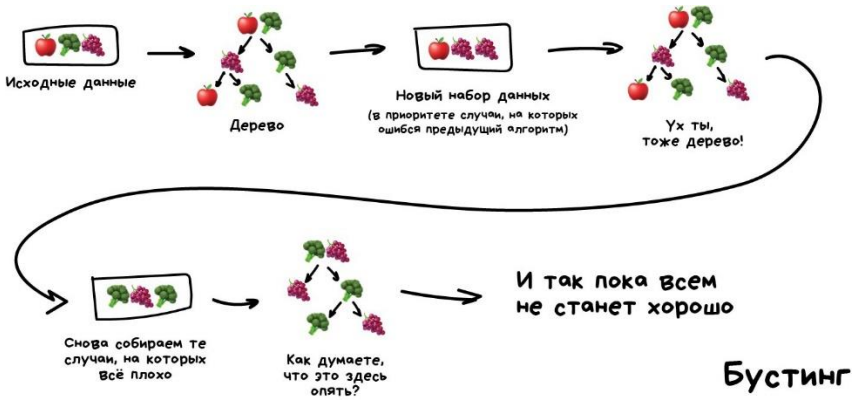
Стекинг работает следующим образом. Обучаются несколько разных алгоритмов и передаются их результаты на вход последнему, который принимает итоговое решение. в качестве решающего алгоритма чаще всего берётся регрессия.

В беггинге обучается один алгоритм много раз на случайных выборках из исходных данных. В самом конце усредняются ответы.



В бустинге обучаются алгоритмы последовательно, каждый следующий уделяет особое внимание тем случаям, на которых ошибся предыдущий. Как в беггинге, делаются выборки из исходных данных, но теперь не совсем случайно. В каждой новой выборке берётся часть тех данных, на которых предыдущий алгоритм отработал неправильно. То есть новый алгоритм как бы доучивается на ошибках предыдущего. Плюсом является

очень высокая точность классификации. Данный способ конструирования ансамблей используется в алгоритме Яндекса матрикснет, который позднее будет описан подробнее.



Регрессия – та же классификация, только вместо категории предсказывается число. Регрессия делится на линейную и полиномиальную.

2.2 Машинное обучение для задачи ранжирования

В ранжировании, как и в любой задаче машинного обучения с учителем, строится функция, которая наилучшим образом соответствует экспертным данным. Экспертами определяется порядок, в котором нужно показывать документы по конкретным запросам. Таких запросов десятки тысяч. И чем лучше, с точки зрения экспертных оценок, оказался порядок документов, сформированный формулой, тем лучшее ранжирование мы получили.

Входными данными для обучаемой функции, по которым она должна определить порядок документов для любого другого запроса, используются так называемые факторы – различные признаки страниц. Эти признаки могут зависеть от запроса (учитывать, сколько его слов содержится в тексте страницы), отличать стартовую страницу сайта от внутренних, использовать признаки самого запроса, которые едины для всех страниц – к примеру, на каком языке задан запрос, сколько в нём слов, насколько часто его задают пользователи и так далее.

Однако, даже собрав достаточное число оценок и рассчитав для каждой пары (запрос + документ) набор факторов, построить ранжирующую функцию стандартными методами оптимизации не так просто. Для облегчения этой задачи и используется обучение ранжированию при помощи машинного обучения.

Во время обучения ранжирующей модели и при её работе, каждая пара документ + запрос переводится в числовой вектор из ранжирующих признаков (также называемых

ранжирующими факторами или сигналами), характеризующих свойства документа, запроса и их взаимоотношение. Такие признаки можно разделить на три группы:

- Статические признаки — зависящие только от документа, но не от запроса. Например, PageRank или длина документа.
- Признаки, зависящие только от запроса.
- Динамические признаки — зависящие и от документа, и от запроса. Например, мера TF-IDF соответствия документа запросу.

Обучающая выборка используется для того, чтобы установить зависимость между порядком страниц для запроса, полученным исходя из их оценки людьми, и признаками этих страниц. Полученная функция используется для ранжирования по всем запросам, независимо от наличия по ним экспертных оценок. В следующем разделе подробно разбираются алгоритмы обучения ранжирования.

2.3 Классификация алгоритмов обучения ранжированию

В зависимости от используемого входного представления данных и функции штрафа алгоритмы обучения ранжированию делятся на три подхода: поточечный, попарный и списочный.

В поточечном подходе предполагается, что каждой паре запрос-документ поставлена в соответствие численная оценка. Задача обучения ранжированию сводится к построению регрессии: для каждой отдельной пары запрос-документ необходимо предсказать её оценку. В рамках этого подхода могут применяться многие алгоритмы машинного обучения для задач регрессии. Когда оценки могут принимать лишь несколько значений, также могут использоваться алгоритмы для ординальной регрессии и классификации. Недостатком поточечного подхода, во-первых, является то, алгоритм рассматривает документы из разных запросов вместе, сравнивая между собой. Например, один человек предпочитает классику, а другой – рок. Зачем определять силу их предпочтения? Во-вторых, любой (правильный и неправильный) порядок объектов с приближенно равными рангами штрафуется функционалом качества одинаково. Так как штраф зависит от величины ранга, а не от порядка.

В попарном подходе обучение ранжированию сводится к построению бинарного классификатора, которому на вход поступают два документа, соответствующих одному и тому же запросу, и требуется определить, какой из них лучше. К нему относятся

следующие алгоритмы: RankNet, FRank, RankBoost, RankSVM, IR-SVM. Недостатками этого подхода является следующее:

- Оптимизируемый фунционал качества оценивает глобальный порядок, а не порядок для одной группы.
- Не учитываются зависимости между сравниваемыми парами в общей группе

Списочный подход заключается в построении модели, на вход которой поступают сразу все документы, соответствующие запросу, а на выходе получается их перестановка. Подгонка параметров модели осуществляется для прямой максимизации одной из перечисленных выше метрик ранжирования. Но это часто затруднительно, так как метрики ранжирования обычно не непрерывны и не дифференцируемых относительно параметров ранжирующей модели, поэтому прибегают к максимизации неких их приближений или нижних оценок. К нему относятся следующие алгоритмы: SoftRank, SVMmap, AdaRank, RankGP, ListNet, ListMLE.

2.4 Модели, используемые в поисковых системах

Поисковые движки многих современных поисковых систем по Интернету, среди которых Яндекс, Yahoo и Bing, используют ранжирующие модели, построенные методами машинного обучения. Поиск Bing использует алгоритм RankNet. Яндекс использует алгоритм собственной разработки MatrixNet, который устойчив к переобучению.

Рассмотрим его подробнее.

С помощью MatrixNet можно построить очень длинную и сложную формулу ранжирования, которая учитывает множество различных факторов и их комбинаций. Другие методы машинного обучения позволяют либо строить более простые формулы с меньшим количеством факторов, либо нуждаются в большей обучающей выборке. Матрикснет строит формулу с десятками тысяч коэффициентов. Это позволяет сделать существенно более точный поиск.

Ещё одна важная особенность MatrixNet — в том, что формулу ранжирования можно настраивать отдельно для достаточно узких классов запросов. Например, улучшить качество поиска только по запросам про музыку. При этом ранжирование по остальным классам запросов не ухудшится. Как же производится ранжирование при помощи MatrixNet?

Поскольку поисковая система работает с очень большими объёмами информации, по каждому запросу ей нужно проверить признаки миллионов страниц, определить их релевантность и соответственно упорядочить. MatrixNet позволяет проверить очень много факторов за короткое время и без существенного увеличения вычислительных мощностей. Поиск ведётся одновременно на тысячах серверов. Каждый сервер ищет по своей части индекса и формирует список самых лучших результатов. В него гарантированно попадают все самые релевантные запросу страницы.

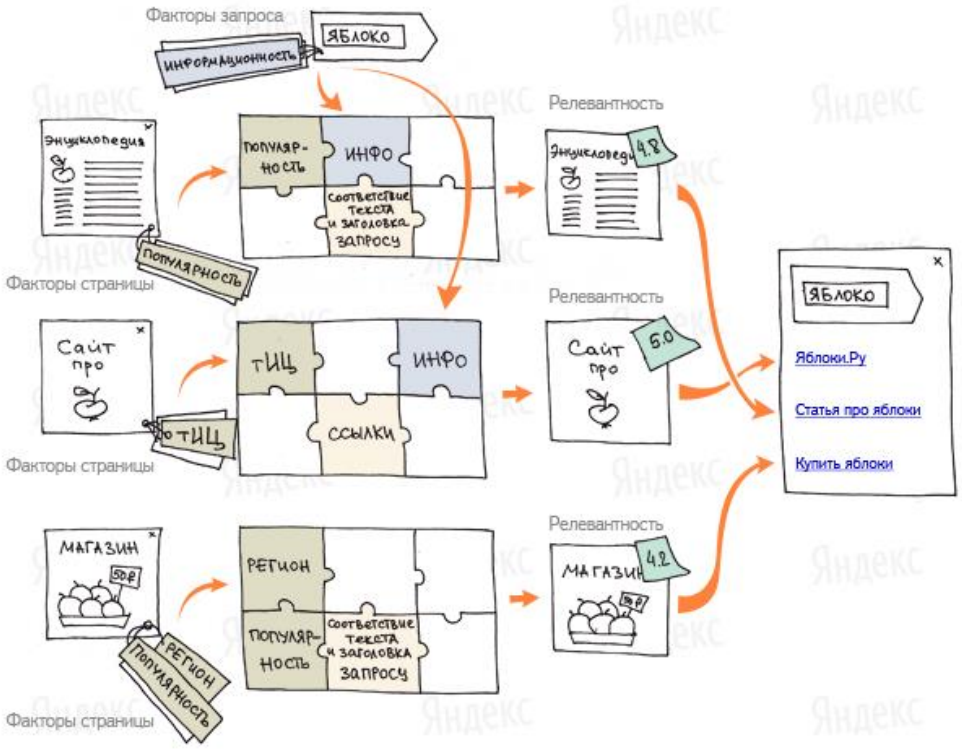


Рис. 2.1. Работа MatrixNet

Дальше из этих списков составляется один общий, и страницы, попавшие туда, упорядочиваются по формуле ранжирования — той самой длинной и сложной формуле, построенной с помощью MatrixNet, с учётом всех факторов и их комбинаций. Таким образом, наверху поисковой выдачи оказываются все самые релевантные сайты – и пользователь почти мгновенно получает ответ на свой вопрос.

ЗАКЛЮЧЕНИЕ

В результате выполнения курсовой работы была изучена задача ранжирования в информационном поиске и то, как машинное обучение помогает в её решении. Были разобраны некоторые алгоритмы машинного обучения, используемые в данной задаче, классификация алгоритмов обучения ранжированию и рассмотрен алгоритм Матрикснет, разработанный компанией Яндекс.

Таким образом, машинное обучение существенно повышает качество решения задачи ранжирования, так как позволяет облегчить построение формулы для ранжирования, используя множество различных факторов, условий и большие объёмы информации.

СПИСОК ЛИТЕРАТУРЫ

1. Маннинг К. Д., Рагхаван П., Шютце Х.: Введение в информационный поиск, ООО “И. Д. Вильямс” – 528 с., 2011;
2. https://habr.com/ru/company/yandex/blog/174213/
3. https://ru.wikipedia.org/wiki/Обучение_ранжированию
4. https://yandex.ru/company/technologies/matrixnet/
5. https://yandex.ru/company/technologies/learning
6. https://vas3k.ru/blog/machine_learning/