Абросимова Н.

ИУ5-41

**Обучение ранжированию**

Для начала следует рассмотреть, что из себя представляет ранжирование.

Ранжи́рование — это сортировка сайтов в поисковой выдаче, применяемая в поисковых системах. Существует множество факторов для ранжирования, среди которых можно отметить рейтинг сайта, количество и качество внешних ссылок, релевантность текста к поисковому запросу, на основании которых поисковая система формирует список сайтов в поисковой выдаче. Алгоритм ранжирования того или иного поисковика меняется в процессе его функционирования

А обучение ранжированию (англ. learning to rank или machine-learned ranking, MLR) — это класс задач машинного обучения с учителем, заключающихся в автоматическом подборе ранжирующей модели по обучающей выборке, состоящей из множества списков и заданных частичных порядков на элементах внутри каждого списка. Частичный порядок обычно задаётся путём указания оценки для каждого элемента (например, «релевантен» или «не релевантен»; возможно использование и более, чем двух градаций).

Применительно к поисковым системам, каждый список представляет собой набор документов, удовлетворяющих некоторому поисковому запросу. Цель ранжирующей модели — наилучшим образом приблизить и обобщить способ ранжирования в обучающей выборке на новые данные. Ведь на каждый запрос может найтись уйма ответов, и далеко не все из них удовлетворяют тому, что действительно нужно пользователю. Поэтому поисковой системе уже недостаточно просто показать все страницы со словами из запроса, она должна расположить найденные страницы в нужном порядке — так, чтобы сверху оказались наиболее подходящие пользователю (наиболее релевантные).

Поисковая система должна научиться строить правило, которое определяет для каждого запроса, какая страница является хорошим ответом на него, а какая — нет. Для этого поисковая машина анализирует свойства веб-страниц и поисковых запросов. Это может быть подготовлено как вручную, специально натренированными людьми (оценщиками качества поиска или асессорами), так и автоматически, на основе анализа пользовательских кликов.

Ранжирующими признаками, из которых состоит числовой вектор, считаются:

* Запросо-независимые или статические признаки — зависящие только от документа, но не от запроса.
* Признаки, зависящие только от запроса
* Запросо-зависимые или динамические признаки — зависящие и от документа, и от запроса.

Кроме факторов ранжирования поисковой системе необходимы образцы — запросы и страницы, которые люди считают подходящими ответами на эти запросы. Оценкой того, насколько та или иная страница подходит для ответа на тот или иной запрос, занимаются специалисты — асессоры, оценивающие, насколько хорошо найденный документ отвечает на заданный запрос.

Так же существует несколько метрик, по которым оценивают и сравнивают качество работы алгоритмов ранжирования на выборке с асессорными оценками:

* DCG и NDCG;
* Точность (precision);
* MAP;
* Mean reciprocal rank;
* pfound — разработка компании Яндекс.

Выделяют три вида алгоритмов для решения задачи обучения ранжированию.

*Поточечный подход*

В поточечном подходе (англ. pointwise approach) 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.

*Попарный подход*

В попарном подходе (англ. pairwise approach) 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.

*Списочный подход*

Списочный подход (англ. listwise approach) 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 недифференцируемы относительно параметров ранжирующей модели, поэтому прибегают к максимизации неких их приближений или нижних оценок.

Обучение ранжированию — это ещё довольно молодая,но бурно развивающаяся область исследований, возникшая в 2000-е годы с появлением интереса в области информационного поиска к применению методов машинного обучения к задачам ранжирования.