Машинное обучение, осень 2016. Никита Волков

Ранжирование

Семинар 2

Ключевые слова

Типы признаков, TF-IDF, BM25, PageRank

1. Признаки пар запрос-документ

Для начала кратко напомним задачу ранжирования. Пусть D — некоторая коллекция текстовых документов и Q — множество запросов. Обозначим D_q неупорядоченный набор документов, потенциально релевантных запросу $q \in Q$. Основная задача ранжирования — упорядочить документы внутри D_q по убыванию степени их релевантности запросу, то есть более релевантные документы должны иметь более высокий ранг.

Оценку релевантности $y(q, d)$ пары запрос-документ (q, d) определяют асессоры. Асессорская выборка X — множество пар (d, q), для которых определена оценка релевантности $y(q, d)$. Таким образом, объектами в данной задаче машинного обучения являются пары запрос-документ (q, d). Однако, все рассмотренные нами методы предполагают наличие их векторного описания. Поэтому нам нужно построить признаки (фичи, факторы) в виде некоторых функций от пар запрос-документ. Обычно их делят на три основных типа

1. зависят только от документа;
2. зависят только от запроса;
3. зависят как от запроса, так и от документа.

2. Текстовые признаки

Текстовые признаки существенно используют тот факт, что как запросы, так и документы являются последовательностью слов.

Пусть у нас есть документ d и запрос q, состоящий только из одного слова $q = (w)$.

Какой самый простой признак можно придумать?

Очедидно, самый простой — индикатор того, что слово w есть в документе d, то есть $f(q, d) = I\{w \in d\}$.

В чем его недостаток и как можно его улучшить?

Слово w может встречаться в разных документах разное количество раз. Давайте будем считать, сколько раз слово w встретилось в документе d, получив признак $f(q, d) = n_{dw} = \sum_{w' \in d} I\{w = w'\}$.

Что здесь плохого и как это можно исправить?

Документы могут быть разной длины. Как следствие, из того, что в документе d слово w было употреблено 3 раза, можно сделать разные выводы о релевантности документа по

запросу в зависимости от длины документа. Можно посчитать частоту слова w в документе d, то есть $f(q,d) = \frac{n_{dw}}{n_d}$, где n_d — количество слов в документе d. Полученная величина называется term frequency и обозначается $TF(w,d)$

Что делать, если в запросе несколько слов?

Просуммируем TF по всем словам из запроса $f(w,d) = \sum_{w \in q} TF(w,d)$.

Некоторые слова могут употребляться сильно чаще других. Что делать?

Есть некоторые общеупотребимые слова, например, предлоги и союзы. Эти слова называются стоп-словами. Можно создать их список, и не учитывать все слова из этого списка.

Однако, так мы только частично решим проблему, ведь все остальные слова так же будут иметь равный приоритет. Вспомним о нашей коллекции документов D. Давайте посчитаем document frequency — частоту, с которой данное слово встречается в документах из коллекции D. Она равна $DF(w) = \frac{|D_w|}{|D|}$, где D_w — множество документов, в которых слово w встречается хотя бы раз. Очевидно, что чем больше значение $DF(w)$, тем менее значимо слово w среди всех слов по запросу q. Тогда в качестве "меры хорошести" слова можно взять обратную частоту, причем обычно еще берут логарифм: $IDF(w) = \log \frac{|D|}{|D_w|}$.

Скрещивая две эти величины, получаем функцию TF-IDF для пары запрос-документ

$$f(q,d) = \text{TF-IDF}(q,d) = \sum_{w \in q} TF(w,d) IDF(w).$$

BM25

Она основывается на вероятностной модели, разработанной в 1970-х и 1980-х годах Стивеном Робертсоном, Карен Спарк Джоунс и другими. Часто называют "Okapi BM25", по названию поисковой системы Okapi, созданной в Лондонском городском университете в 1980-х и 1990-х годах, в которой эта функция была впервые применена.

$$BM25(q,d) = \sum_{w \in q} IDF(w) \frac{TF(w,d)(k+1)}{TF(w,d) + k\left(1 - b + b\frac{n_d}{Al}\right)},$$

где n_d — число слов в документе d, Al — среднее число слов в документе по всем документам из коллекции, $k \in \mathbb{R}_+$ и $b \in [0,1]$ — параметры. Обычно считают, что $k = 2$, $b = 0.75$.

При $k = 0$ формула соответствует простому IDF. При $k \to +\infty$ и $b = 0$ формула соответствует TF-IDF.

Описание и другие варианты http://kak.tx0.org/IR/TFxIDF

3. Примеры других признаков

- Домен сайта. Например, $f(q,d) = I\{\text{сайт находится в зоне .ru}\}$. Особенность этого признака в том, что он не только не зависит от запроса, но еще и является хостовым, то есть принимает одно значение для всех документов данного сайта.
- Авторитетный сайт. Например, для Википедии $f(q,d) = I\{d \text{ — страница с Википедии}\}$.
- Входит в какую-то определенную категорию сайтов. Например, новостной сайт или сервис ответов.

- Язык сайта. Например, $f(q, d) = I\{\text{в } d \text{ большинство слов русские}\}$. Или же можно считать долю русских слов.
- Размер документа. Можно понимать в разных смыслах — количество слов, размер занимаемой памяти.
- Наличие на странице особых объектов, например, изображение, музыка, видео.
- Язык запроса.
- Длина запроса.
- Запрос содержит вопрос.

3. Признаки на основе PageRank

HostRank
Аналогичен PageRank, но не для отдельных страниц, а для сайтов в целом.

TrustRank
Недобросовестные вебмастера стремятся захватить верхние строчки поисковой выдачи, продвинув туда свои зачастую не очень качественные и не представляющие ценности для пользователей ресурсы. Описание метода TrustRank для решения задачи отделения полезных сайтов от бесполезных впервые сделали специалисты поисковика Yahoo.

В некоторой мере он подобен PageRank, за исключением того, что по ссылкам с одних сайтов на другие передаётся не вес, а некий уровень доверия. Начальный набор хороших страниц и хороших сайтов задается вручную, экспертами, оценивающими качество сайтов. Значение зависит от положительных факторов (авторитетные сайты, корпорации, каталоги) и отрицательных факторов (скрытый текст, клоакинг, редиректы)