Министерство образования и науки Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего профессионального образования

«Московский физико-технический институт (государственный университет)»

Факультет инноваций и высоких технологий

Кафедра анализа данных

На правах рукописи
УДК ????

Левков Мирон Николаевич

Гладкая метрика для задачи ранжирования

Выпускная квалификационная работа бакалавра

Направление подготовки: 010400 Прикладные математика и информатика

Заведующий кафедрой	_____	/???/
Научный руководитель	_____	/Воронцов А.?./
Студент	_____	/Левков М.Н./

г. Москва
2017

Гладкая метрика для задачи ранжирования

М.Н. Левков

Аннотация. Рассматривается задача ранжирования документов в поисковых запросах. В работе исследованы имеющиеся варианты сглаживания метрики ранжирования. Предложены различные способы сглаживания метрики DCG. Проведен анализ зависимости качества сглаженных метрик от гиперпараметров и размера выборки поисковых запросов.

Содержание

1 Введение . . . 2
2 Метрики качества результатов . . . 3
3 Рассмотренные метрики . . . 5
3.1 Метрика SoftDCG . . . 5
3.2 Метрика NoisedSoftDCG . . . 6
3.3 Метрика FairSoftDCG . . . 7
4 Применение результатов . . . 8
5 Эксперименты . . . 9
5.1 Зависимость гладкости от размера пула запросов . . . 9
5.2 Зависимость качества аппроксимации от размера пула запросов . . . 10
5.3 Изменение метрики при добавлении шума . . . 11
6 Проблема выбора гиперпараметра . . . 12
7 Выводы . . . 13

Список литературы . . . 14

Введение

Рассмотрим задачу ранжирования документов в выдаче поиского запроса.

Обозначение 1.1. Обозначим $\{d_i\}_{i=1}^n$ - набор *документов* релевантных данному поисковому запросу; $\{r_i\}_{i=1}^n$ - *рельные оценки* данных документов, проставленные ассесорами; $\{s_i\}_{i=1}^n$ - оценки релевантности (*скоры*), выданные ранжирующим алгоритмом

Определение 1.1. Метрика качества ранжирования **DCG** определяется по формуле:

$$DCG = \sum_{i=1}^{k} \frac{r_{p_i}}{discount(i)}$$

где $p_1, ..., p_n$ - перестановка на множестве $\{1, ..., n\}$, т.ч. $s_{p_1} > s_{p_2} > ... > s_{p_n}$; k - количество документов, по которым считается метрика ($k \leq n$); discount(i) - дисконтирующий фактор, как правило $\frac{1}{i}$

Традиционные метрики ранжирования имеют конструкцию похожую на DCG: несложно заметить, что при фиксированном наборе документов и их оценок данная метрика принимает конечное количество различных значений. При этом изменение значения метрики происходит лишь в случае перестановки местами двух документов в выдаче. Таким образом метрика не имеет гладкой зависимости от скоров. В то же самое время логично ожидать, что, если ранжирующий алгоритм выдал трем документам оценки $\{100, 1, 0.5\}$, а другим трем документам - $\{10, 1, 0.5\}$, то он считает первый документ из первой тройки сильно лучшим, чем первый документ из второй тройки. В данной работе исследуются разные подходы к построению метрик ранжирования, с целью получения метрики, которая бы удовлетворяла ряду свойств, вводимых далее

Метрики качества результатов

Хотелось бы, чтобы наша метрика обладала двумя свойствами. Во-первых, метрика должна быть "реалистичной". Т.е. ее локальные минимумы/максимумы должны соответствовать минимумам/максимумам DCG.

Определение 2.1. Пусть $\{y_i\}_{i=1}^k$ - значения метрики DCG на данном пуле запросов для k различных значений набора гиперпараметров. Пусть $\{\hat{y}_i\}_{i=1}^k$ - значения нашей метрики на том же пуле при тех же наборах гиперпараметров. Тогда **качеством аппроксимации** метрики DCG нашей метрикой назовем величину $\min\limits_{\alpha,\beta} \frac{1}{k} \sum\limits_{i=1}^{k} (y_i - \hat{y}_i)^2$

Обозначение 2.1. Далее будем обозначать *качество аппроксимации*, как approx(Metric Name)

Также придуманная метрика должна быть гладкой. При этом необходим способ подсчета гладкости конкретной функции, если она задана не аналитически, а численно. Определим метрик гладкости.

Определение 2.2. Пусть $\{y_i\}_{i=1}^k$ - значения нашей метрики на данном пуле запросов для k различных значений набора гиперпараметров. Тогда **гладкостью** функции, посчитанной **через усреднение модуля разности** в соседних точках, будем называть величину

$$\text{smooth}_{\text{abs}}(y_1, ..., y_k) = \sum_{i=2}^{k} |y_i - y_{i-1}| \cdot \frac{1}{|y_k - y_1|}$$

Определение 2.3. Пусть $\{y_i\}_{i=1}^k$ - значения нашей метрики на данном пуле запросов для k различных значений набора гиперпараметров. Тогда **гладкостью** функции, посчитанной **через нормировку дисперсии** разности в соседних точках, будем называть величину

$$\text{smooth}_{\text{std}}(y_1, ..., y_k) = \frac{diff_{std}}{|diff_{mean}|}$$

Здесь $diff_{mean} = \frac{1}{k-1} \sum\limits_{i=2}^{k} (y_i - y_{i-1}), \quad diff_{std} = \frac{1}{k-1} \sum\limits_{i=2}^{k} (y_i - y_{i-1} - diff_{mean})^2$

Определение 2.4. Пусть $\{y_i\}_{i=1}^k$ - значения нашей метрики на данном пуле запросов для k различных значений набора гиперпараметров. Пусть $window \in \mathbb{Z}, deg \in \mathbb{N}$. Тогда **гладкостью** функции, посчитанной **с помощью аппроксимации полиномами**, будем называть величину

$$\text{smooth}_{\text{poly}}(y_1, ..., y_k) = \frac{1}{k - window} \sum_{i=1}^{k-window+1} \left(poly_{deg}(y_i, ..., y_{i+window-1})_{i+\lfloor \frac{window}{2} \rfloor} - y_{i+\lfloor \frac{window}{2} \rfloor} \right)^2$$

Здесь $poly_{deg}(y_i, ..., y_{i+window-1})$ - аппроксимирующий полином степени deg, построенный по window точкам. Т.е. такой полином стпепени deg, что среднеквадратичное отклонение в точках $\{i, ..., i + window - 1\}$ минимально.

В терминах данных нами определений, новая метрика тем лучше, чем меньше такие показатели, как гладкость и аппроксимация.

Возникает ряд проблем с тем, что выбор конкретного способа измерения гладкости для сравнения наших метрик между собой неочевиден. Первая метрика хорошо отображает гладкость в том случае, если функция монотонна (если не учитывать шумовые колебания) - однако же в иных случаях данная метрика может быть плоха из-за нормировки на разность значений в крайних точках.

Вторая метрика ведет себя лучше, однако тоже не является достаточно гибкой и интерпритируемой.

С последней метрикой встает проблема выбора параметров deg и window. Тем не менее эта метрика понятна и действительно способона достаточно хорошо отображать гладкость функции. После перебора разных вариантов выбор был сделан в пользу параметров $deg = 4$, $window = 11$. При этом стоит учитывать, что гладкость метрики ранжирования измерялась на множествах размера порядка 100 точек.

Рассмотренные метрики

Метрика SoftDCG

Метрика NoisedSoftDCG

Метрика FairSoftDCG

Применение результатов

Эксперименты

Зависимость гладкости от размера пула запросов

Зависимость качества аппроксимации от размера пула запросов

Изменение метрики при добавлении шума

Проблема выбора гиперпараметра

Выводы

Список литературы