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

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

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