

Гладкие метрики в задаче ранжирования

Научрук: Александр Воронцов

Студент: Мирон Левков

25 июня 2017 г.

Проблема

У метрик ранжирования есть одна проблема - метрика меняется лишь при перестановке двух документов местами. Таким образом метрики не имеют гладкой зависимости от скоров, выдаваемых ранжирующим алгоритмом.

Постановка задачи

$$DCG = \sum_{i=1}^p \frac{rel_i}{discount(i)}$$

Целью работы является поиск метрики ранжирования, которая удовлетворяет следующим свойствам:

- Данная метрика является достаточно гладкой (о том, как измерялась гладкость, далее)
- Данная метрика достаточно хорошо линейно аппроксимирует DCG, т.е. $DCG \simeq \alpha \cdot metric + \beta$

Применение результатов

Гладкий аналог метрики DCG является более чувствительным способом измерения качества модели.

- Смешивание моделей
- Отбор признаков
- Кривая обучения модели
- Добавление шума в качестве sanity-check

Метрики качества результатов

- 1 Гладкость кривой (значения метрики с изменением некоего параметра - например, коэффициента смеси двух формул) определялась так:

$$\frac{1}{n - window} \sum_{i=1}^{n-window} (polydeg(y_i, \dots, y_{i+window-1})_{i+\frac{window}{2}} - y_{i+\frac{window}{2}})^2$$

- 2 Аппроксимация метрики DCG считалась, так:

$$\min_{\alpha, \beta} MSE(values_{dgc}, \alpha \cdot values_{metric} + \beta)$$

Рассмотренные метрики. SoftDCG

- 1 Сопоставим каждому документу нормальное распределения с некими параметрами.
- 2 Посчитаем мат ожидание позиции в выдаче для каждого документа, чтоб использовать средний discount [1]

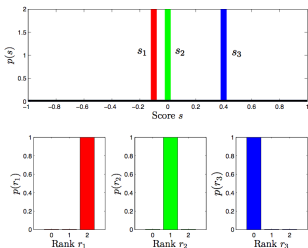


Figure 1: Deterministic scores and ranks: Three document scores as point (deterministic) values and their corresponding rank distributions. The lowest scoring document s_1 is certain to be ranked in the lowest position 2.

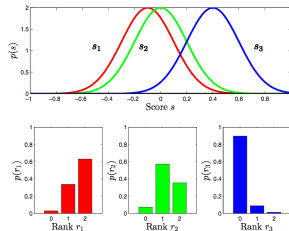


Figure 2: From score to rank distributions: Smoothed scores for 3 documents and the resulting 3 rank distributions.

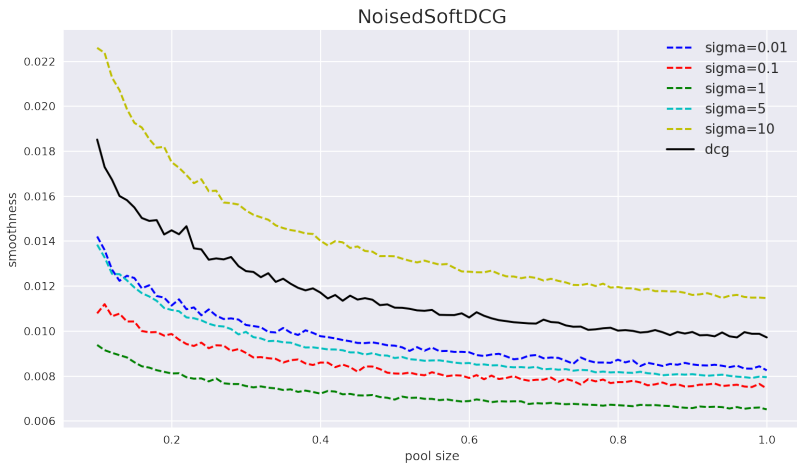
Рассмотренные метрики. NoisedSoftDCG

- 1 Сгенерируем случайный шум и добавим его к скорам
- 2 Посчитаем средний DCG по m зашумлениям
- 3 Шум может менять между собой близкие по скорам документы тем самым сглаживая метрику при усреднении

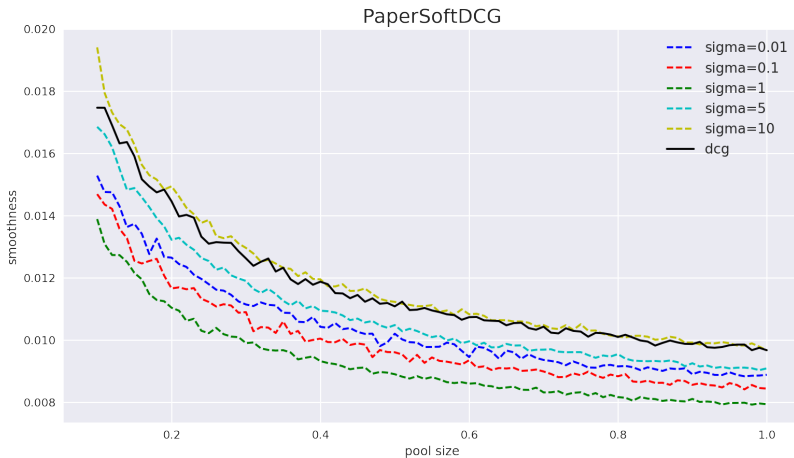
Рассмотренные метрики. FairSoftDCG

- 1 Будем интерпретировать скоры, как вероятности документов находиться на своих позициях. Т.е. $P(doc_i > doc_j) = \frac{e^{\sigma s_i}}{e^{\sigma s_i} + e^{\sigma s_j}}$
- 2 Можно ввести распределение на всех перестановках документов
- 3 Посчитаем мат ожидание DCG по всем перестановкам документов в top-k

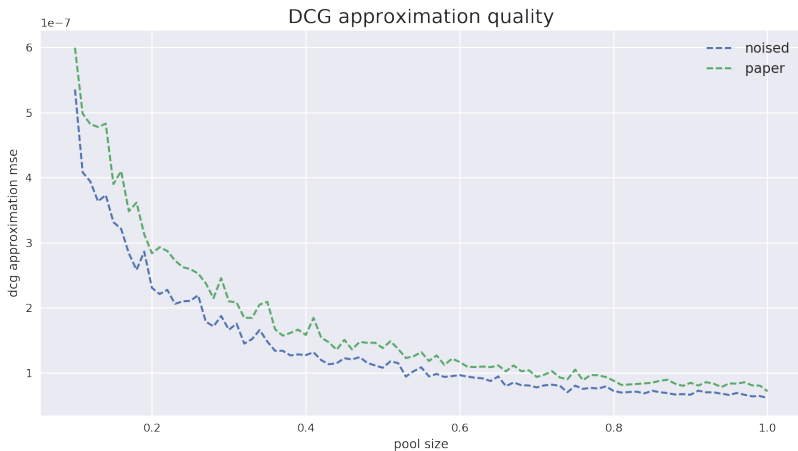
Гладкость метрик при изменении размера пула



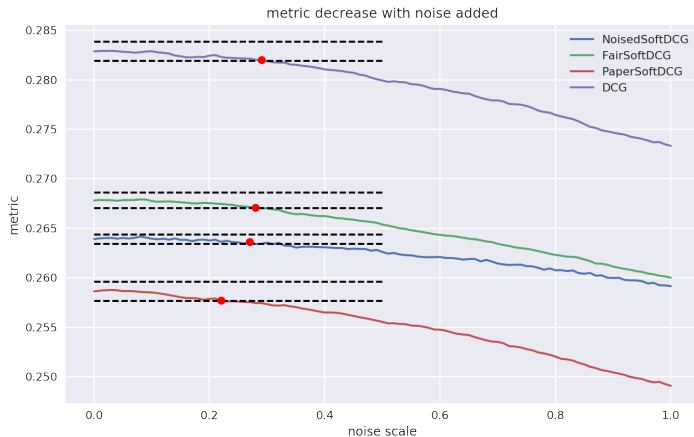
Гладкость метрик при изменении размера пула



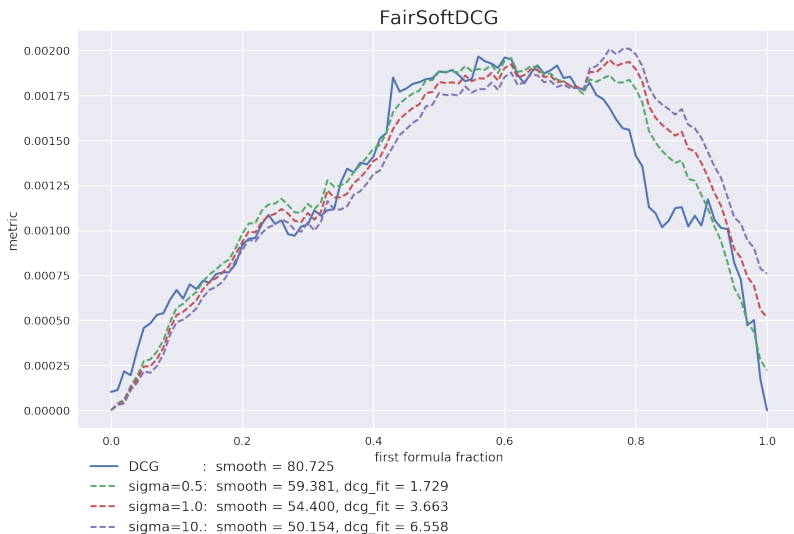
Аппроксимация DCG при изменении размера пула



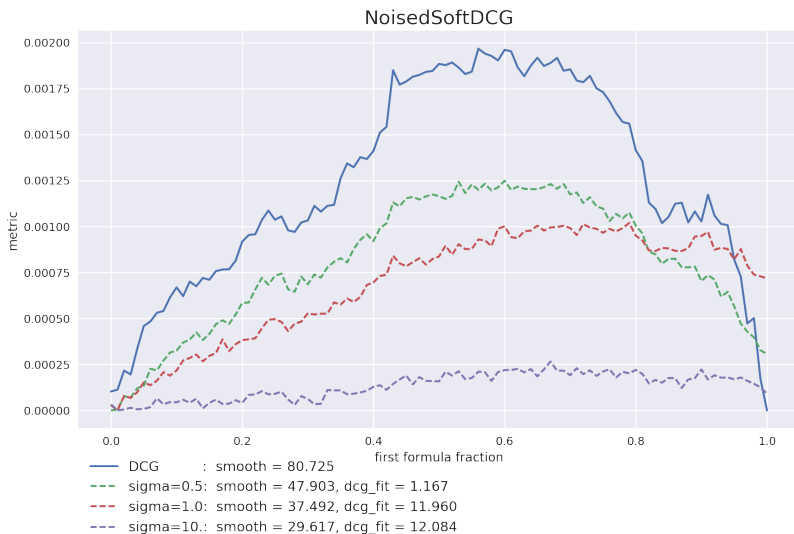
Изменение метрик при зашумлении



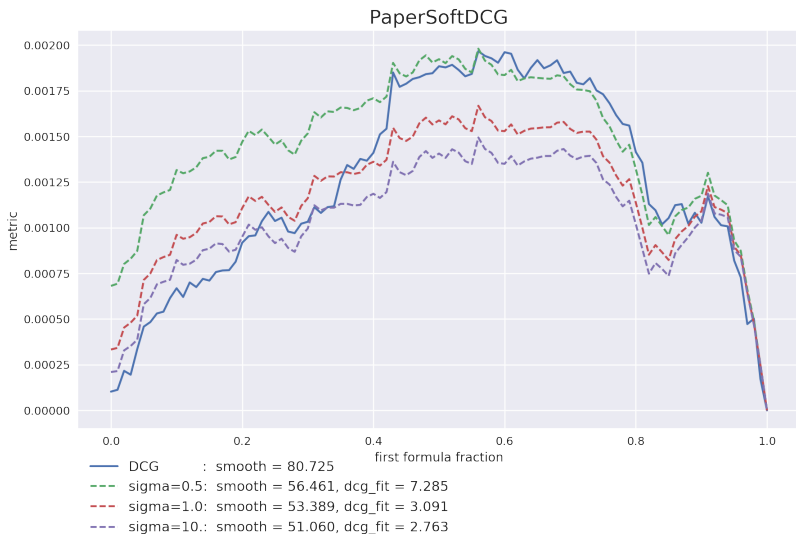
Метрика для выпуклой комбинации формул (max/min)



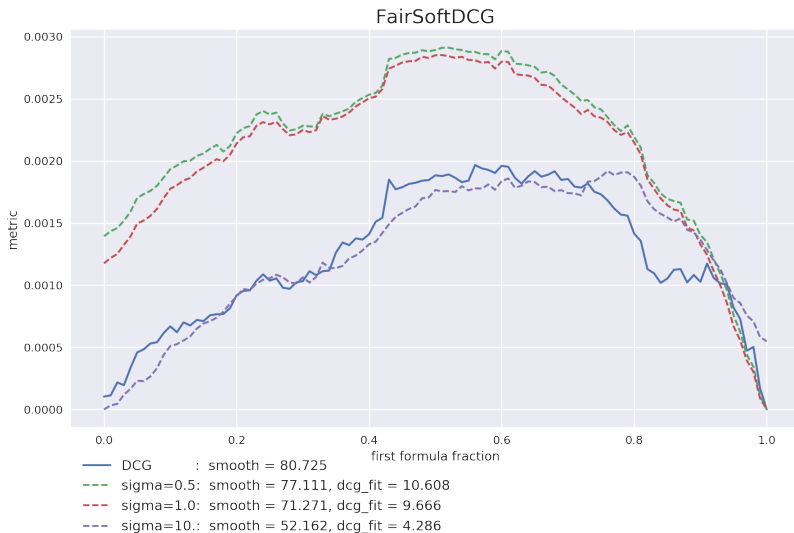
Метрика для выпуклой комбинации формул (max/min)



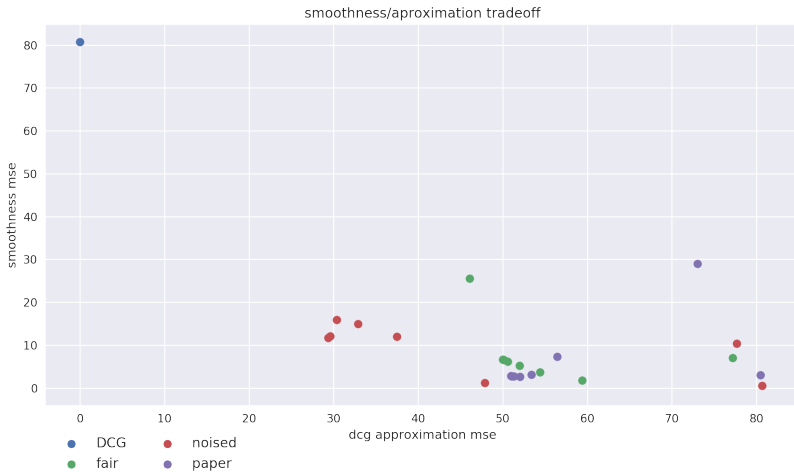
Метрика для выпуклой комбинации формул (max/min)



Метрика для выпуклой комбинации формул (std)



Выбор сигмы



Проблемы

- ❶ Хочется выбирать параметр σ нашей метрики автоматически
- ❷ Наша метрика гладкости не является идеальной, т.к. тоже требует подбора параметров: степень аппроксимирующего полинома, размер окна

Список литературы

- [1] M. Taylor, J. Guiver, S. Robertson and T. Minka. SoftRank: Optimising Non-Smooth Rank Metrics. Microsoft Research Cambridge, 2016
- [2] Christopher J.C. Burges. From RankNet to LambdaRank to LambdaMART. Microsoft Research Technical Report, 2010
- [3] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, Hang Li. Learning to Rank: From Pairwise Approach to Listwise Approach. Microsoft Research Technical Report, 2007

Спасибо за внимание!