NDCG – Normalized Discounted Cumultative Gain

Pedro Henrique Silva Rodrigues
Bacharelado em Ciência da Computação
IFG – Instituto Federal de Goiás Câmpus Anápolis
28 de agosto de 2018

Como identificar se uma predição ou ranqueamento realmente são bons e estão de acordo com a classificação do usuário?

Medidas de Ranqueamento

Os algoritmos de ranqueamento nos informam uma lista de recomendação ou classificação com base em parâmetros pré-determinados. Ou seja retornam uma lista de itens ranqueados. Para saber se esse ranking gerado foi bom utilizamos medidas de ranqueamento para avaliá-las.

Medidas de Ranqueamento

Algumas medidas de ranqueamento muito utilizadas:

- MAP (Mean Average Precision)
- NDCG (Normalized Discounted Cumulative Gain)
- MRR (Mean Reciprocal Rank)

DCG

É o ganho acumulado até a posição k de uma coleção k de docuementos ranqueados.

Na posição i, com (i>2), é o somatório das relevâncias a partir do segundo documento, até a posição i, descontado a posição:

Se i = 1

$$DCG(i) = rel(i)$$

Se i > 1

$$DCG(i) = DCG(i - 1) + rel(i) / \log(i + 1)$$

DCG

- DCG é a soma acumulada da relevância de itens ranqueados.

Doc	Relevance	i
d1	2	1
d2	3	2
d3	2	3
d4	4	4

$$DCG = 2 + 3/\log(2) + 2/\log(3) + 4/\log(4)$$

IDCG

É o ganho acumulado até a posição k de uma coleção k de docuementos ranqueados, porém considerando o caso ótimo.

Na posição i, com (i>2), é o somatório das relevâncias dos valores ordenados a partir do segundo documento, até a posição i, descontado a posição.

IDCG

- Já o NDCG é a soma acumulada da relevância de itens ranqueados.

Doc	Relevância	i	Relevância Ótima
d1	2	1	4
d2	3	2	3
d3	2	3	2
d4	4	4	2

$$IDCG = 4 + 3/\log(2) + 2/\log(3) + 2/\log(4)$$

NDCG

Por fim, o NDCG é o ganho acumulado normalizado, ou seja, a razão do DCG obtido pelo IDCG (melhor caso).

- Trata-se portanto de uma escala que varia de 0 a 1. Onde quanto mais próximo de um mais próximo está da resposta esperada.

NDCG

Para o exemplo anterior ficaria:

DCG = 2 + 3/log(2) + 2/log(3) + 4/log(4)
= 2 + 3/1 + 2 / 2.0959 + 4/2
= 7.9542

IDCG = 4 + 3/log(2) + 2/log(3) + 2/log(4)
= 4 + 3/1 + 2/2.0959 + 1
= 8.9542

NDCG = DCG / IDCG = 7.9542 / 8.9542

NDCG = 0.8883

Referências:

- Hamed Valizadegan, Rong Jin, Ruofei Zhang, and Jianchang Mao. 2009. Learning to rank by optimizing NDCG measure. In Proceedings of the 22nd International Conference on Neural Information Processing Systems (NIPS'09), Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Eds.). Curran Associates Inc., USA, 1883-1891.

- LEE, Huei Diana. Seleção e construção de features relevantes para o aprendizado de máquina. 2000. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2000. doi:10.11606/D.55.2000.tde-15032002-113112. Acesso em: 2018-09-02.