CUMULATIVE GAIN, DISCOUNTED CUMULATIVE GAIN, NORMALIZED DCG

Overview—

Discounted cumulative gain (DCG) is a measure of ranking quality. In information retrieval, it is often used to measure effectiveness of web search engine algorithms or related applications. Using a graded relevance scale of documents in a search-engine result set, DCG measures the usefulness, or gain, of a document based on its position in the result list. The gain is accumulated from the top of the result list to the bottom, with the gain of each result discounted at lower ranks.

Two assumptions are made in using DCG and its related measures.

- Highly relevant documents are more useful when appearing earlier in a search engine result list
- Highly relevant documents are more useful than marginally relevant documents, which are in turn more useful than non-relevant documents

Cumulative Gain—

Cumulative Gain (CG) is the preprocessor of DCG and does not include the position of a result in the consideration of the usefulness of a result set. In this way, it is the sum of the graded relevance values of all results in a search result list. The CG at a particular rank position p is defined as:

$CG_p = \sum_{i=1}^{p} rel_i$

Where rel_i is the graded relevance of the result at position i

The value computed with the CG function is unaffected by changes in the ordering of search results. That is, moving a highly relevant document d_i above a higher ranked, less relevant, document d_j does not change the computed value for CG. Based on the two assumptions made above about the usefulness of search results, DCG is used in place of CG for a more accurate measure.

Discounted Cumulative Gain—

The premise of DCG is that highly relevant documents appearing lower in a search result list should be penalized as the graded relevance value is reduced logarithmically proportional to the position of the result. The traditional formula of DCG accumulated at a particular rank position p is defined as:

$$DCG_p = \sum_{i=1}^{p} \frac{rel_i}{log_2(i+1)} = rel_i + \sum_{i=2}^{p} \frac{rel_i}{log_2 i+1}$$

An alternative formulation of DCG places stronger emphasis on retrieving relevant documents:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i}-1}{log_2 i+1}$$

This form is commonly used in industry including major web search companies and data science competition platform such as Kaggle.

Normalized DCG—

Search result lists vary in length depending on the query. Comparing a search engine's performance from one query to the next cannot be consistently achieved using DCG alone, so the cumulative gain at each position for a chosen value of p should be normalized across queries. This is done by sorting all relevant documents in the corpus by their relative relevance, producing the maximum possible DCG through position p, also called Ideal DCG (IDCG) through that position. For a query, the normalized discounted cumulative gain, or nDCG, is computed as:

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

where IDCG is ideal discounted cumulative gain,

$$IDCG_p = \sum_{i=1}^{|REL|} \frac{2^{rel_i}-1}{log_2 i+1}$$

and $|REL|$ represents the list of relevant documents (ordered by their relevance) in the corpus up to position p.