Proposal for Evaluation of GenIR System Using NQC

Varun Bharti, Shubham Kale

July 2024

1 Executive Summary

This proposal outlines a novel approach to evaluate generative systems for relevance retrieval when manually assessed relevant documents (rel docs) are unavailable. Our objective is to leverage the Normalized Query Commitment (NQC) method, a standard Query Performance Prediction (QPP) technique, as a substitute for rel docs. This approach allows us to assess the effectiveness of generative systems in producing relevant responses to user queries.The anticipated outcome is a documented approach for evaluating generative systems using NQC-based estimated relevance, providing valuable insights into their performance without relying on manual assessments.

2 Overview

The evaluation of generative systems for relevance retrieval traditionally relies on manually assessed rel docs. These documents, judged by human experts to be relevant to a specific query, serve as a benchmark for evaluating the system's ability to generate relevant responses. However, obtaining manually assessed rel docs can be time-consuming, expensive, and sometimes infeasible. This proposal addresses this challenge by proposing an alternative evaluation approach using NQC. This work contributes to the development of more efficient and adaptable evaluation methods for generative models.

3 Proposed Methodology

3.1 Data and Resources:

- **Retrieval System:** We will need both standard retrieval methods like BM25, ColBERT-H etc. along with Generative models. This system should be able to generate responses to a given query.
- **Standard Datasets:** We propose using established datasets : TREC DL 2019 , TREC DL 2020 and MS MARCO V1 collection.

3.2 Retrieval and NQC Calculation:

- Instead of directly comparing generated responses with judged relevant passages, we'll estimate relevance scores using NQC.
- For each query in the TREC dataset:
  - Execute the query using the chosen retrieval model (e.g., BM25) on the preprocessed document collection.
  - Retrieve the top k documents (e.g., k = 20) based on their relevance scores assigned by the retrieval model.
  - Calculate the NQC score for the retrieved documents using the following steps:
    - Compute the standard deviation (SD) of the retrieved document scores.
    - Divide the SD by the average score (μ) of the retrieved documents (NQC = SD / μ).
- Estimated Relevance Mapping:
  - Map the NQC value for each retrieved document to an estimated relevance score between 1 and k (assuming k is the chosen top document limit) using following manner :
    - Calculate the relative fraction of NQC within its possible range (0 to 1) by dividing the NQC value by the maximum possible SD (which equals the average score, μ).
    - Multiply the obtained fraction by k (the target range for the mapped value).
    - Round the resulting value to the nearest integer to get the estimated relevance score for that document.

3.3 Evaluation with Estimated Relevance:

- Generate responses for each query using generative system. Instead of comparing generated responses with the actual judged relevant passages, compare them with the retrieved documents with the highest estimated relevance scores (based on NQC mapping).
- Analyze the correlation between the embedding similarity scores of generated responses and the estimated relevance scores obtained from NQC mapping.
- A high correlation suggests that NQC effectively captures the relevance distribution within the retrieved documents and the generated responses are aligning well with those considered highly relevant by the NQC estimation.

3.4 Basic Key Differences Between our Approach and Original Approach :

Feature	**Original Approach**	**Our Appraoch(NQC)**
Relevance Source	Manually judged relevant documents	NQC-based estimated relevance
Similarity Comparison	Judged relevant passages	Retrieved documents with highest estimated relevance
Performance Metric	Embedding similarity to judged passages	Embedding similarity to estimated relevance

4 References:

- *Negar Arabzadeh, Amin Bigdeli, and Charles L. A. Clarke*: Adapting Standard Retrieval Benchmarks to Evaluate Generated Answers , Link to Paper
- *Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber and Gad Markovits*: Predicting Query Performance by Query-Drift Estimation, Link to Paper
- *Suchana Datta, Debasis Ganguly, Mandar Mitra, Derek Greene*:A Relative Information Gain-based Query Performance Prediction Framework with Generated Query Variants, Link to Paper
- *Haggai Roitman* : Normalized Query Commitment RevisitedLink to Paper