Fréchet Distance for Offline Evaluation of Information Retrieval Systems with Sparse Labels

Anonymous EACL submission

Abstract

The rapid advancement of natural language processing, information retrieval (IR), computer vision, and other technologies has presented significant challenges in evaluating the performance of these systems. One of the main challenges is the scarcity of human-labeled data, which hinders the fair and accurate assessment of these systems. In this work, we specifically focus on evaluating IR systems with sparse labels, taking inspiration from the success of using Fréchet Inception Distance (FID) in assessing text-to-image generation systems. We propose leveraging the Fréchet Distance to measure the distance between the distributions of relevant judged items and retrieved results. Our experimental results on MS MARCO V1 dataset and TREC Deep Learning Tracks query sets demonstrate the effectiveness of the Fréchet Distance as a metric for evaluating IR systems, particularly in settings where a few labels are available. This approach contributes to the advancement of evaluation methodologies in real-world scenarios such as the assessment of generative IR systems.

1 Introduction

With the rapid advancement of technologies in fields such as natural language processing, natural language generation, computer vision, and information retrieval (IR), evaluating the performance of these systems is becoming increasingly challenging (Gatt and Krahmer, 2018; Hashimoto et al., 2019; Celikyilmaz et al., 2020; Yang and Lerch, 2020). We must develop new metrics, benchmarks, and evaluation protocols that are specifically tailored to the unique characteristics of the systems considering the rapid changes in system architecture, training data, and model configurations (Theis et al., 2015). In many cases, obtaining high-quality labeled data that accurately represents the complexity of real-world scenarios can be expensive, time-consuming, or even impractical. This scarcity of labeled data adds to the limitations of conducting extensive evaluations and may lead to biased or incomplete assessments (Arabzadeh et al., 2022).

Offline evaluation poses a significant challenge due to the sparsity of labeled data (Clarke et al., 2023, 2020; Xie et al., 2020). This challenge is particularly prominent in datasets like MS MARCO, a widely used benchmark for ad hoc retrieval reserach (Nguyen et al., 2016; Arabzadeh et al., 2021; Mackenzie et al., 2021) in which, the majority of queries are annotated with only one relevant judged document. However, to suit the dataset for effective traininig of deep learning models, a high number of queries are judged, resulting in sparse labels. Consequently, most queries have only one relevant judgment, while the relevance of the remaining documents remains unknown. Other researchers have shown that there are potentially relevant documents that are as good as, or even better than, the judged queries (Qu et al., 2020; Arabzadeh et al., 2022). Given the sparsity of ground truth labels, it is crucial to recognize the challenges involved in distinguishing between rankers when the differences in performance are small (Yan et al., 2022). The limited labeled data for retrieved documents introduces noise, making it challenging to definitively determine which ranker is performing better (Cai et al., 2022). The incomplete judgments can introduce problems in evaluations, as they do not capture the full range of relevant documents (Aslam et al., 2006; Carterette and Smucker, 2007). This issue becomes even more pronounced in generative-based tasks. It is impractical to reassess the generated results, such as images or text, with each system run due to their non-deterministic nature (Theis et al., 2015; Harshvardhan et al., 2020).

Evaluating a generative system's performance based on the similarity of generated content to sparsely labeled data remains one of the most effective approaches in many generative-based NLP and computer vision benchmarks and tasks (Soloveitchik et al., 2021; Heusel et al., 2017; Obukhov and Krasnyanskiy, 2020; Dimitrakopoulos et al., 2020; Zhang et al., 2019). Particularly in the evaluation of text-to-image generation task, the Fréchet Inception Distance (FID), has gained recognition for showing high robustness and correlation with human judgements (Heusel et al., 2017; Saharia et al., 2022; Yu et al., 2022). FID compares the distribution of generated images across a set of prompts to the distribution of target images across the same set of prompts. To compute FID, features of ground truth images and generated images are extracted from both sets, and multivariate Gaussian distributions are fitted to these features. The Fréchet Distance (*FD*), which quantifies the similarity between two probability distributions, is then computed based on the fitted Gaussian distributions. A lower FID score indicates a higher similarity between the distributions, indicating that the generated images closely match the real images in terms of their visual features.

In this paper, we shed light on how evaluating generated results is similar to assessing the quality of retrieved results with sparse labels in an ad hoc retrieval setting. Most benchmarks for both tasks have quite sparse labels i.e., not all the items are judged and while there are a few annotations available for some of the candidates, there can be other unjudged relevant items available. While labelling more data is expensive for both tasks, there could be more than one correct answer in both tasks. In this work, we mimic an Information Retrieval system with sparse relevance judgements as a generation task where the ground truth targets are sparse. Due to the success of FID in evaluating the quality of generated images, especially for generative adversarial networks (Gafni et al., 2022; Saharia et al., 2022; Yu et al., 2022; Khan et al., 2020; Alonso et al., 2019), we explore if we can quantify the quality of retrieved documents in an ad hoc retrieval system through *Fréchet Distance*. In the context of IR evaluation, we can analogously consider the relevant judged items as the ground truth set and the retrieved items as the set of generated items. Our objective is to extract features from both sets, the relevant judged items and the retrieved results, and investigate whether metrics such as the Fréchet Distance can effectively capture the quality of the retrieved results with respect to the ground truth labels in IR systems.

We study the following Research Questions:

- RQ1. Can the Fréchet Distance effectively evaluate IR systems with sparse labels?
- RQ2. Can the Fréchet Distance effectively evaluate IR systems with comprehensive labels?
- RQ3. Can the Fréchet Distance effectively evaluate the quality of IR systems when the retrieved results are not labelled?
- RQ4. How well correlated are the performance of IR systems, as measured by the Fréchet Distance vs. and traditional IR metrics?
- RQ5. How robust is the Fréchet Distance for evaluating IR systems with respect to the feature extraction methods used to represent both the ground truth and retrieved items?

We conduct our experiments by assessing different retrieval pipelines on the MS MARCO V1 Dev dataset, which has extremely sparse labels, as well as the TREC Deep Learning Track 2019 and 2020 datasets, which have more complete labels (Nguyen et al., 2016; Craswell et al., 2020, 2021). Our study demonstrates the effectiveness of the Fréchet Distance as a metric for quantifying the performance of IR systems especially when the ground truth labels are sparse.

2 Fréchet Distance for IR evalaution

2.1 Fréchet Distance

The Fréchet distance is a measure of dissimilarity between two curves or trajectories and has shown to be useful in numerous applications including computational geometry, computer graphics, bioinformatics and robotics (Alt, 2009; Alt and Godau, 1995; Alt et al., 2001; Jiang et al., 2008; Gheibi et al., 2014). To understand the Fréchet distance, let us consider two curves (or trajectories or paths): A and B. The Fréchet distance between A and B could be exemplified as measuring the minimum leash length required by a dog walking along a path A while its owner walks along path B, with both the dog and owner potentially traversing their respective paths at different speeds (Alt and Buchin, 2007; Eiter and Mannila, 1994). The leash cannot be shortened or lengthened during the walk. The definition is symmetric i.e., the Fréchet distance would be the same if the dog were walking its owner. Given two curves, A and B, represented as sequences of points in a metric space, the Fréchet distance, denoted as $F(A, B)$ is computed as:

$$F(A,B) = inf_{\alpha,\beta} max_{t\in[0,1]} d(A(\alpha(t)), B(\beta(t))) \quad (1)$$

where A and B are continues maps from $[0, 1]$ to metric space and α and β are reparameterizations of the unit interval $[0, 1]$ i.e. they are continuous, non-decreasing, surjection functions. The requirement of non-decreasing reparameterizations, α and β, ensures that neither the dog nor its owner can backtrack along their respective curves. The parameter t as represents the progression of time, consecutively $A(\alpha(t))$ and and $B(\beta(t))$ represent the position of the dog and the dog's owner at time t (or vice versa). The distance d between $A(\alpha(t))$ and $B(\beta(t))$ corresponds to the length of the leash between them at time t. By considering the *infimum* over all potential reparameterizations of the unit interval $[0, 1]$, we select the specific paths where the maximum leash length is minimized.

Apart from quantifying the dissimilarity between curves, the Fréchet distance can also serve as a measure to assess the disparity between probability distributions (Heusel et al., 2017).Given we have two normal univariate distributions, X and Y, Fréchet Distance (*FD*) is given as:

$$FD(X, Y) = (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2 \tag{2}$$

Where μ and σ are the mean and standard deviation of the normal distributions, respectively.

2.2 Fréchet Inception Distance

In computer vision, the Inception V3 model pretrained on the Imagenet dataset is employed to generate feature vectors to be approximated by multivariate normal distribution (Szegedy et al., 2015). As such, the Fréchet Inception Distance (FID) for a multivariate normal distribution is computed as:

$$FID(X, Y) = ||\mu_X - \mu_Y||^2 - Tr(\Sigma_X + \Sigma_Y - 2\sqrt{\Sigma_X \Sigma_Y}) \tag{3}$$

In this equation, X and Y represent two distributions derived from two sets of embeddings. These embeddings correspond to real images and generated images, respectively, and are obtained from the Inception model. The vectors X and Y have magnitudes μ_X and μ_Y, respectively. The trace of the matrix is denoted as Tr, while Σ_X and Σ_Y represent the covariance matrices of the vectors.

2.3 Fréchet Distance for IR

Let us assume C is a collection of items and $Q = \{q_1, q_2, \ldots, q_n\}$ is a set of n queries, where each query q_i has a set of relevant judged items R_{q_i}. We define R_Q as a set of relevance judged items for queries in Q, where $R_Q = \{d | d \in R_{q_i}, q_i \in Q\}$. Furthermore, we can obtain the top-k retrieved items by a retrieval system M from C for a given query q as $M_k(q, C) = D_q^k$, where D_q^k is a set of the top-k most relevant retrieved items for query q, i.e., $D_q^k = \{d_1^q, d_2^q, \ldots, d_1^k\}$. Given $\mathbb{V}$ as a function that maps any retrieved item to a p-dimensional embedding space, we can embed all the retrieved items and relevant judged items through $\mathbb{V}$. For instance, $\mathbb{V}(d_1)$ returns a p-dimensional vector embedding for document d_1. To apply Fréchet Distance for assessing the quality of the IR system M, we measure FD_Q^M as follows on query set Q:

$$FD_Q^{M_k} = FD\Big(\{\mathbb{V}(R_Q)\}, \{\mathbb{V}(M_k(Q, C))\}\Big) \tag{4}$$

Here, *FD* is the Fréchet Distance (Eq. 3) measures the distance between the distribution of the set embeddings of the relevant judged items $\{\mathbb{V}(R_Q)\}$ and those of the retrieved items $\{\mathbb{V}(M_k(Q, C))\}$. The lower $FD_Q^{M_k}$ represents the retrieved items to have higher similarity with the relevant judged items and thus the better performance of the retrieval system M on the query set Q.

3 Experimental Setup

3.1 Dataset and Query sets

We perform experiments on the MS MARCO passage retrieval collection V1 , which includes over 8.8 million passages (Nguyen et al., 2016). First, in section 4, we experiment on the 6980 queries in MS MARCO small dev set, which are sparsely labelled. The majority of the queries in this set (over 94%) have only one relevant judged document per query. Second, in Section 5, we experiment on the TREC Deep Learning (DL) track 2019 and 2020 to study how varying and extending the relevance judgments would affect the evaluation process (Craswell et al., 2021, 2020). The difference between the two query sets is that while the MS MARCO dev set has a higher number of queries (6980) judged, with mostly one relevant document per query, it leaves us with no extra information about the unannotated documents. On the other hand, the TREC DL tracks have fewer queries judged (97), but each query has a comprehensive set of judgments with multi-level judgments ranging from 0-4, indicating the degree of relevance.

We compare the results of the *FD* score with the official traditional IR evaluation metrics of each benchmark, i.e., MRR@10 for MS MARCO and nDCG@10 for TREC Deep Learning tracks.

3.2 Retrieval models

We consider a set of 12 retrieval methods that are well-distinguished for their efficiency or effectiveness, ranging from traditional high-dimensional bag-of-word sparse retrievers to more recent dense retrievers well as trained high-dimensional sparse models.Specifically, we consider BM25 as the representative of the sparse retrievers standalone as well as applying BM25 to expanded documents through DeepCT and DocT5Query document expansion methods (Robertson et al., 1995; Nogueira et al., 2019a,b; Dai and Callan, 2019). We include a set of dense retrievers including RepBERT (Zhan et al., 2020), ANCE (Xiong et al., 2020), Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), COLBERT (Khattab and Zaharia, 2020) and COLBERT-V2 (Santhanam et al., 2021). We also employ the more recently proposed high dimensional learnt sparse retrievers, UniCOIL and SPLADE (Formal et al., 2021; Lin and Ma, 2021). Furthermore, we consider hybrid retrievers (Lin et al., 2021b) that fuse the retrieved items from BM25 and dense retrievers, to cover a variety of retrievers and assess the ability of *FD* to quantify the quality of retrieval fairly. We note that we employ some of the retrieval models from Pyserini(Lin et al., 2021a) and some of the others from the paper's original GitHub repository. For more information about each of the retrieval models, we kindly refer to the original papers of each method.

For our experiments with the TREC DL19 and DL20 query sets, we took the submitted runs for each track from the NIST website[1]. Our experiments compare the results when assessing with Fréchet distance as well as nDCG@10 for 37 submitted runs to TREC DL2019 and 59 submitted runs to TREC DL 2020. These runs cover a comprehensive set of retrieval pipelines, typically with from sparse and/or dense retrieval as a retrieval first stage followed by one or more neural re-ranking stages (Craswell et al., 2020, 2021).

3.3 Embeddings

To examine the robustness of *FD* on IR systems, we perform experiments using two different types of transformer-based contextualized models to embed the documents and extract their features. We employ a general-purpose DistilBERT (Sanh et al., 2019) to obtain the documents embeddings[2] as well as fine-tuned pre-trained language models on MS MARCO[3] (Reimers and Gurevych, 2019). Both models were adapted from hugging face. We note that unless we explicitly mention (Section 7.2) all the results are reported with the first model, i.e., the DistilBERT model that was fine-tuned on MS MARCO. We believe that by exploring different document representations, we may better understand the influence of document quality on the utilization of *FD* for evaluating IR systems.

Table 1: Performance of different retrievers in terms of MRR@10 as well as Fréchet distance *FD* on MS MARCO dev set. A smallest Fréchet distance corresponds to better performance.

Category	Method	MRR@10	*FD*@1	*FD*@10
Sparse	BM25	0.187	7.446	4.410
	DeepCT	0.242	1.453	2.354
	DocT5	0.276	3.047	2.050
Dense	RepBERT	0.297	1.881	1.223
	ANCE	0.330	1.529	0.995
	SBERT	0.333	1.387	1.008
	ColBERT	0.335	1.456	0.980
	ColBERT V2	0.344	1.453	0.982
Trained Sparse	UniCOIL	0.351	1.387	0.980
	SPLADE	0.368	1.328	0.964
Hybrid (BM25)	ColBERT-H	0.353	1.494	0.973
	ColBERT V2 -H	0.368	1.464	0.998

4 Assessment with Sparse labels

We are interested in investigating how *FD* can assess the performance of different retrievers when there are only sparse labels available i.e., on 6980 queries from MS MARCO small dev set. We present the performance of the 12 retrieval methodsintroduced in Section 3.2 in terms of MRR@10 as well as measuring the Fréchet Distance between two sets of retrieved items and relevant judged items on the cut-offs of 1 and 10 in Table 1.

The results for *FD*@1 and *FD*@10 demonstrate the ability of *FD* to quantify the performance of retrievers. For example, for the BM25 retriever, *FD*@1 is measured as 7.446 and *FD*@10 as 4.410. However, for a neural retriever like ColBERT, which has shown superior performance to BM25 on various benchmarks (Santhanam et al., 2021; Khattab and Zaharia, 2020; Thakur et al., 2021), the *FD* values are reported as 1.456 and 0.980 for *FD*@1 and *FD*@10, respectively. This indicates that *FD* can effectively pickout the *better* retriever, particularly when there is a significant difference between their performances. On the other hand,

[1]https://trec.nist.gov/

[2]https://huggingface.co/distilbert-base-uncased

[3]https://huggingface.co/sentence-transformers/msmarco-distilbert-base-v2

when the performance of two retrievers is quite similar, such as in the case of ColBERT vs. ColBERT V2, it becomes more challenging for evaluation metrics to assess their performance . For instance, while MRR@10 for ColBERT vs. ColBERT V2 is reported as 0.334 vs. 0.343, *FD*@10 for the two retrievers is reported as 0.980 and 0.982. Therefore, as expected, the discriminative power of *FD* decreases when it becomes harder to distinguish between retrievers. However, It is important to acknowledge that the noise introduced by limited labeled data for retrieved documents makes it difficult to definitively determine which ranker is performing better (Qu et al., 2020). In fact Arabzadeh et al. (2022) showed that such a small difference in MRR@10 is not a strong indicator of which retrieval method is able to address the queries better since they might have surfaced other *unjudged relevant items*. They showed that ordering of the rankers solely based on MRR and incomplete relevance judgement is not reliable. Based on the results in Table 1 and their comparison with MRR@10, we can conclude that in response to **RQ1**, we observe that Fréchet Distance can effectively evaluate IR systems.

To examine the robustness of the *FD* in the context of IR assessment, and to evaluate the generalizability of the method across different subsets of queries, we employ a bootstrap sampling (Johnson, 2001; Efron, 2003) from the MSMARCO dev set for $N = 1000$ times. This would allows us to investigate whether the results obtained in the previous section were influenced by the data or if they can be reliable. The results are visualized in Figure 1, in which we present the mean and empirical 0.95% confidence interval for each retriever across the 1000 query sets in terms of MRR@10 and *FD*@10. It is important to note that for the MRR plot, a higher position on the plot indicates better performance, while for the *FD* plot, a lower position indicates better performance. The findings confirm that despite considering different sample sets, we observe a consistent pattern and similarity in the performance trends.

5 Assessing with Comprehensive labels

In this section, we investigate the performance of the Fréchet Distance in evaluating IR systems when the labels are not sparse and we have more complete labels. We conduct experiments using the runs submitted to TREC DL 2019 (37 runs)

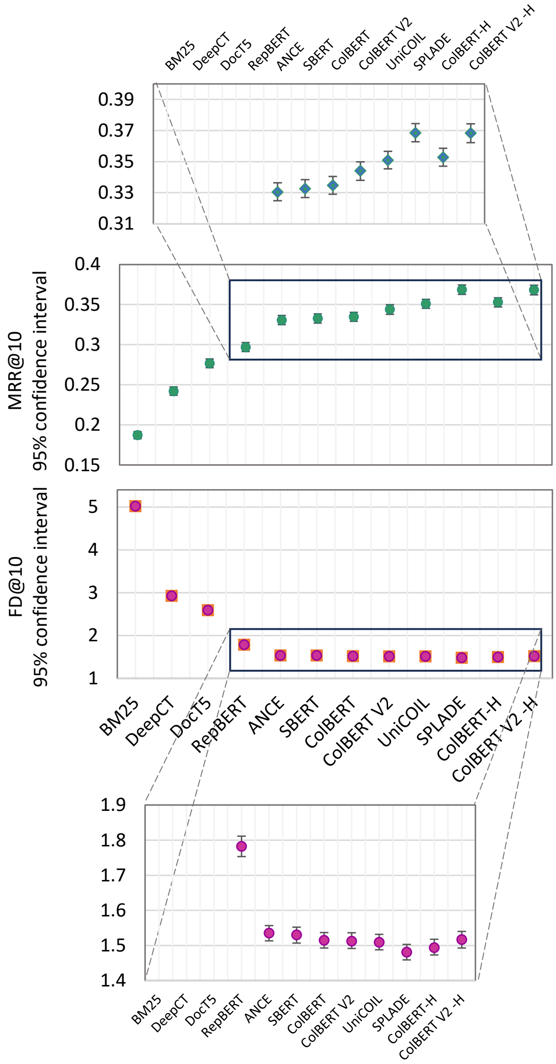

Figure 1: Performance of bootstrap sampling (N=1000) of queries in MS MARCO dev set in terms of MRR@10 and *FD*@10 for the 12 different retrieval methods.

and TREC DL 2020 (59 runs). Unlike the MS MARCO dev set which on average each query has 1.06 judged documents, the queries in TREC DL tracks on average have over 210 judged documents per query assessed with four different levels of relevanceCraswell et al. (2020). We notice that the number of judged relevant items per query in these benchmarks varies a lot. Due to the TREC-style judgment criteria, only the top few retrieved items from all submitted runs were judged. Depending on the overlap between the top retrieved items from different runs, the number of relevant judged items per query may vary. When applying *FD* with an imbalanced number of relevant judged items per query, it can introduce biases in the ground truth distribution and potentially lead to problems in evaluation. To address this issue, we balanced the number of relevant judged items per query by limiting

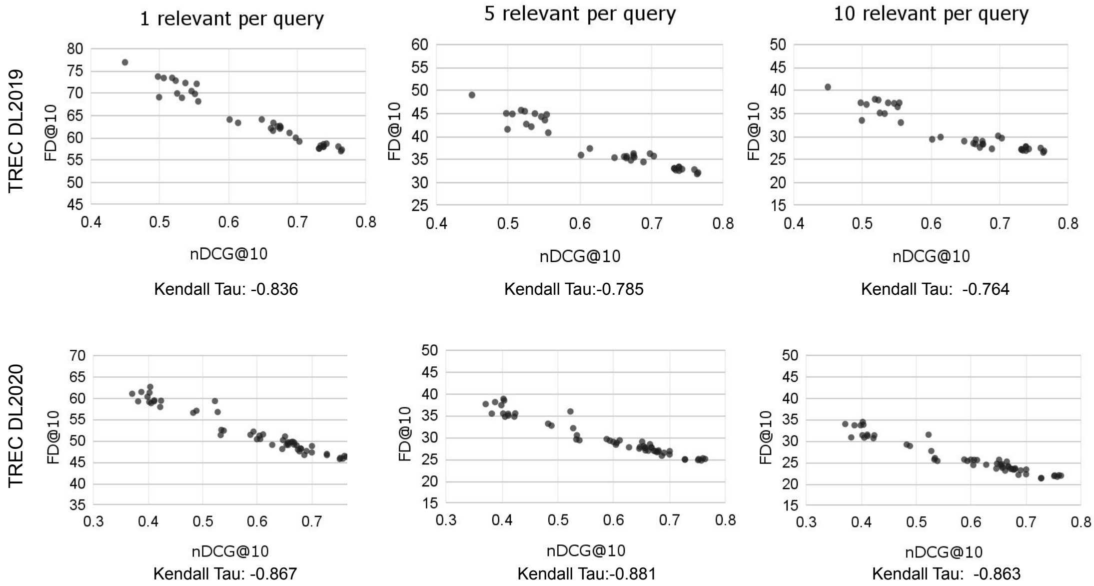

Figure 2: Performance of all the submitted runs to TREC DL 2019 (first row) and TREC DL 2020 (second row). In each sub-figure, X-axis and Y-axis indicate nDCG@10 and *FD*@10 respectively. *FD*@10 was measured with 1,5 and 10 relevant items per query in the sub-figures in the first, second and third columns respectively.

them to a maximum of 1, 5, and 10 relevant judged items per query i.e., we randomly select K relevant items from the pool of relevant judged documents for the query of interest. We first randomly select from the most relevant level i.e., level 3 which are perfectly relevant documents and then when there is not a sufficient number of perfectly relevant documents, we move on to highly relevant level and randomly choose from that grade. This experiment also allows us to examine how the sparsification of judgments affects the performance of evaluation metrics. We note that these modifications in relevance judgements are only applied for measuring *FD* and nDCG@10 is measured with all the judged documents without any modification.

We plotted the nDCG@10 on the x-axis and the *FD* with balanced and sparsified judgments on the y-axis of each sub-figure in Figure 2, for all the runs submitted to TREC DL19 (first row) and TREC DL20 (second row). Consistent with our previous experiments, we observe a highly linear relationship between the two metrics. We also provide the Kendall τ correlation under each sub-figure. For instance, when sparsifying the labels and considering only one relevant judged item per query, we obtain a Kendall τ correlation of -0.836 for TREC DL2019 and -0.867 for TREC DL2020, between nDCG@10 and *FD*@10 of each dataset.

The experiments on the TREC DL datasets highlight two key points. First, unlike using the Fréchet Inception Distance to evaluate the quality of generated images in text-to-image generation tasks, where a large number of data points (in the order of thousands) are required for the evaluation to be valid, we demonstrated that even with a smaller number of queries (around 40-50), *FD* is capable of distinguishing the performance of different rankers (Kynkäänniemi et al., 2023; Heusel et al., 2017). Second, *FD* is not sensitive to the sparsity of the ground truth labels and it performs well with both sparse and more complete labels. It is not affected by the number of judgments, as evidenced by the fact that the performance did not differ greatly when increasing the number of relevant judged items. However, for TREC DL2019, we observed a small drop in correlation by increasing the number of relevant judgments. Further exploration revealed that a higher number of relevant judgments in TREC 2019 resulted in a higher usage of level 2 relevance judgments (highlight relevant) instead of level 3 judgments (perfectly relevant). Consequently, we suggest that *FD* may be more sensitive to the quality of relevant judged items rather than the quantity. Overall, in response to **RQ2**, we find that *FD* works well when using comprehensive labels, and consistent with the findings in Section 4, sparsifying the labels does not compromise the quality of assessment.

6 Assessing Unlabeled Retrieved Results

Here, we undertake an evaluation of different IR systems under an extremely challenging case of assessing unlabeled retrieved results. This scenario presents a situation where each query is assumed

Table 2: Performance of different retrievers in terms of MRR@10 as well as Fréchet distance *FD* assuming under Unlabeled Retrieved Results (URR) setting. We note that the MRR@10 is measured on the original ranked list since with URR setting, all the retrievers would obtain MRR@10 equals to zero. A smallest Fréchet distance corresponds to better performance.

			URR	
Category	Method	MRR@10	*FD*@1	*FD*@10
Sparse	BM25	0.187	8.634	4.705
	DeepCT	0.242	4.183	2.591
	DocT5	0.276	4.066	2.290
Dense	RepBERT	0.297	2.701	1.364
	ANCE	0.330	2.353	1.126
	SBERT	0.333	2.266	1.156
	ColBERT	0.335	2.308	1.115
	ColBERT V2	0.344	2.352	1.121
Trained Sparse	UniCOIL	0.351	2.302	1.128
	SPLADE	0.368	2.300	1.117
Hybrid (BM25)	ColBERT-H	0.353	2.399	1.115
	ColBERT V2 -H	0.368	2.365	1.142

to have mostly only one relevant item, and the *relevant judged items are not included in the top-k results*. Our objective is to investigate the effectiveness of the Fréchet Distance in assessing the top-k Unlabeled Retrieved Results (URR) when no judgments are available for any of the top-k retrieved items. This is particularly valuable considering the high cost and limited availability of labeled data, which often exhibit sparsity. Previous research has demonstrated that as rankers improve in performance, they tend to retrieve previously unseen content that may be highly relevant to the original query (Arabzadeh et al., 2022). If Fréchet Distance is capable of evaluating the retrieved results in such cases, it would be a valuable tool for assessing the relevance of unlabeled data and even beyond that, for evaluating generative-based responses.

We measure the *FD* between one set consisting of the relevant judged items per query and the other set consisting of the top-k *unjudged* retrieved item for each query. In other words, we scan down the ranked list and retain the first k unjudged document to assess. This is an interesting aspect to study because traditional IR metrics such as MRR, nDCG, and MAP rely on the presence of relevant items in the retrieved list and would assign a performance score of zero in cases where no relevant items are retrieved. They do not account for unjudged documents. We argue that by utilizing the *FD* metric, we can capture the similarity between unjudged retrieved items and the limited set of judged examples and measure the performance of the retriever based on this value.

The results of this experiment are reported in Table 2 with two cut-offs of "*FD*@10" and "*FD*@1". Even when no judged documents appear in the top-k, *FD* is still able to quantify the performance of the retriever. This capability is not present in traditional metrics. For instance, when there are no relevant judged items retrieved in the ranked list, *FD*@1 quantifies the performance of BM25 as 8.634, whereas the performance for ColBERT is measured as 2.308. This indicates that even without relevant judged items, *FD* is capable of determining that ColBERT performs better than BM25.

Table 3: Kendall τ correlation between different evaluation metrics over the 12 retrieval methods. URR stands for "Unlabeled Retrieved Results" and refers to experimental results from section 6. All the correlations are statistically significant with p-value < 0.05

	MRR@10	*FD*@1	*FD*@1 URR	*FD*@10	*FD*@10 URR
MRR@10	1	-0.473	-0.545	-0.788	-0.636
FD@1	-0.473	1	0.687	0.443	0.290
FD@1-URR	-0.545	0.687	1	0.636	0.485
FD@10	-0.788	0.443	0.636	1	0.848
FD@10-URR	-0.636	0.29	0.485	0.848	1

This experiment demonstrates that, unlike traditional IR metrics, *FD* is not sensitive to the labeled documents themselves. Indeed, the Fréchet Distance is not reliant on the exact positioning of the relevant judged document in the ranking. Instead, it focuses on measuring the similarity between the retrieved items and the relevant judged documents. This characteristic makes it particularly valuable for evaluating scenarios with extremely sparse labels, even in cases where the rankers do not retrieve the labeled data. In response to **RQ3**, the Fréchet Distance enables assessment of the remaining unlabeled data, offering valuable insights into their relevance. *In contrast, traditional IR metrics would be unable to provide any insights without retrieving the labeled documents.*

7 Further analysis

7.1 Correlation with IR Evaluation Metrics

We aim to examine the correlation between the *FD* measure and traditional IR evaluation metrics. To achieve this, we calculate the ranked-based Kendall τ correlation, for each pair of metrics in Table 1 and Table 2 on the performance of the 12 retrievers introduced earlier and report the results in Table 3. This set of evaluation metrics includes MRR@10, *FD* at cut-offs 1 and 10 (Section 4) and *FD* at cut-offs 1 and 10 under URR setting when no labeled data is retrieved (Section 6). As anticipated and illustrated in Figure 2, *FD* exhibits a nega-

Table 4: Comparison of the performance of different retrievers when assessing with MRR@10 and *FD*@10 on MS MARCO dev set With DistilBERT fine-tuned on MSMARCO as well as DistilBERT without any fine-tuning. DistilBERT fine-tuned on MSMARCO shows -0.788 Kendall τ correlation with MRR@10 and DistilBERT without any fine-tuning shows -0.739 Kendall τ correlation with MRR@10.

Category	Method	MRR@10	*FD*@10 DistilBERT MSMARCO	*FD*@10 DistilBERT No Fine-tuning
Sparse	BM25	0.187	0.590	4.410
	DeepCT	0.242	0.412	2.354
	DocT5	0.276	0.331	2.050
Dense	RepBERT	0.297	0.159	1.223
	ANCE	0.330	0.121	0.995
	SBERT	0.333	0.132	1.008
	ColBERT	0.335	0.117	0.980
	ColBERT V2	0.344	0.118	0.982
Trained Sparse	UniCOIL	0.351	0.123	0.980
	SPLADE	0.368	0.120	0.964
Hybrid (BM25)	ColBERT-H	0.353	0.116	0.973
	ColBERT V2 -H	0.368	0.126	0.998

tive correlation with MRR, as a lower *FD* value indicates better performance. Among these correlations, *FD*@10 shows the highest absolute correlation with MRR@10 i.e., a correlation of -0.788. We suggest that this is because *FD* operates based on the distribution of embedded representations of documents, which has shown to work most stably when the number of samples increases (Chong and Forsyth, 2019; Bińkowski et al., 2018). More interestingly, *FD*@1 and *FD*@1 with Unlabeled Retrieved Results (URR), obtain a correlation coefficient of 0.687. Similarly, the correlation between *FD*@10 (Fréchet Distance at 10) and *FD*@10 with unlabeled retrieved items was found to be 0.848. The high correlation between evaluating the original retrieved results vs without having any judged retrieved results further validates the findings presented in sections 4 and 6.The Fréchet Distance not only exhibits a high correlation with traditional IR metrics but also demonstrates its capability in assessing unlabeled retrieved items. These experiments let us answer **RQ4** that *FD* shows a notable correlation with traditional IR metrics. These properties increase the reliability of using *FD* for assessing IR systems.

7.2 Impact of Document Representation

Here, we examine the robustness of the Fréchet Distance metric for assessing IR systems with respect to the underlying language model to embed the retrieved documents and relevance judgments. For previous experiments, we utilized a language model that was fine-tuned on the MS MARCO dataset for ranking tasks. However, now we study how the results would be impacted if we were to embed the retrieved documents and ground truth in a different space. As such, we present the same results as in Table 1, using DistilBERT embeddings fine-tuned on the MSMARCO training set as well as the same results with a DistilBERT without any fine-tuning. This analysis aims to investigate whether a general-purpose language model can capture the necessary information for accurate assessment, or if a language model specifically fine-tuned for ranking tasks in retrieval is required. Table 4 displays the obtained results. Surprisingly, we observe that changing the language model from a fine-tuned ranking model to a raw, unfine-tuned BERT model does not substantially impact the assessment outcomes. The *FD* metric remains capable of effectively evaluating the performance of various retrieval methods. For example, from Table 4, and under "DistilBERT No fine-tuning" column, we observe that BM25 achieves an *FD*@10 score of 4.410, whereas COLBERT, which is expected to be a better model, achieves a score of 0.980.

The correlation between *FD*@10 and MRR@10 when using DistilBERT without any fine-tuning, is -0.739. Comparatively, when using fine-tuned DistilBERT (as shown in Table 3), the correlation IS -0.788. As such, having a fine-tuned language model specifically for ranking task can improve the correlation with traditional IR metrics. However, even without fine-tuning, *FD* still demonstrates promising performance. Overall, the results indicate that *FD* remains effective in evaluating the quality of retrieved results, even when employing a general-purpose language model without fine-tuning. Lastly, with respect to **RQ5**, we note that *FD* shows promising robustness w.r.t the document embedding representation.

8 Conclusion and Future work

In this paper, we leverage Fréchet Distance to address the challenges of evaluating IR systems with sparse labels. Through experiments conducted on datasets with sparse and more complete ground truth labels, we demonstrated that the Fréchet Distance has significant implications for evaluating IR systems in real-world settings where obtaining comprehensive ground truth labels can be challenging and expensive. We believe that future research could utilize the Fréchet Distance to evaluate different generative models, expanding the scope of evaluation in IR systems.

9 Limitations

While our study provides valuable insights into the effectiveness of the Fréchet Distance in evaluating IR systems with sparse labels, there are a few limitations that should be acknowledged. First, unlike traditional IR evaluation metrics, the Fréchet Distance is not applicable to individual queries and can only be used with sets of queries. Further exploration is needed to understand how the sample size of the queries affects the quality of the assessment. Second, the Fréchet Distance assumes that the two distributions follow a multivariate normal distribution. Lastly, it is important to note that the Fréchet Distance is an unbounded metric, and its range varies depending on the dataset's characteristics and the number of samples under investigation. Building upon the findings of this study,

References

Eloi Alonso, Bastien Moysset, and Ronaldo Messina. 2019. Adversarial generation of handwritten text images conditioned on sequences. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 481–486. IEEE.

Helmut Alt. 2009. The computational geometry of comparing shapes. *Efficient Algorithms: Essays Dedicated to Kurt Mehlhorn on the Occasion of His 60th Birthday*, pages 235–248.

Helmut Alt and Maike Buchin. 2007. Can we compute the similarity between surfaces? *CoRR*, abs/cs/0703011.

Helmut Alt and Michael Godau. 1995. Computing the fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 5(01n02):75–91.

Helmut Alt, Christian Knauer, and Carola Wenk. 2001. Matching polygonal curves with respect to the fréchet distance. In *STACS 2001: 18th Annual Symposium on Theoretical Aspects of Computer Science Dresden, Germany, February 15–17, 2001 Proceedings 18*, pages 63–74. Springer.

Negar Arabzadeh, Bhaskar Mitra, and Ebrahim Bagheri. 2021. Ms marco chameleons: challenging the ms marco leaderboard with extremely obstinate queries. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4426–4435.

Negar Arabzadeh, Alexandra Vtyurina, Xinyi Yan, and Charles LA Clarke. 2022. Shallow pooling for sparse labels. *Information Retrieval Journal*, 25(4):365–385.

Javed A Aslam, Virgil Pavlu, and Emine Yilmaz. 2006. A statistical method for system evaluation using incomplete judgments. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 541–548.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Yinqiong Cai, Jiafeng Guo, Yixing Fan, Qingyao Ai, Ruqing Zhang, and Xueqi Cheng. 2022. Hard negatives or false negatives: Correcting pooling bias in training neural ranking models. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 118–127.

Ben Carterette and Mark D Smucker. 2007. Hypothesis testing with incomplete relevance judgments. In *Proceedings of the sixteenth ACM conference on conference on information and knowledge management*, pages 643–652.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Min Jin Chong and David A. Forsyth. 2019. Effectively unbiased FID and inception score and where to find them. *CoRR*, abs/1911.07023.

Charles LA Clarke, Fernando Diaz, and Negar Arabzadeh. 2023. Preference-based offline evaluation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1248–1251.

Charles LA Clarke, Alexandra Vtyurina, and Mark D Smucker. 2020. Offline evaluation without gain. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, pages 185–192.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the TREC 2020 deep learning track. *CoRR*, abs/2102.07662.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the trec 2019 deep learning track. *arXiv preprint arXiv:2003.07820*.

Zhuyun Dai and Jamie Callan. 2019. Context-aware sentence/passage term importance estimation for first stage retrieval. *arXiv preprint arXiv:1910.10687*.

Panagiotis Dimitrakopoulos, Giorgos Sfikas, and Christophoros Nikou. 2020. Wind: Wasserstein inception distance for evaluating generative adversarial network performance. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3182–3186. IEEE.

Bradley Efron. 2003. Second thoughts on the bootstrap. *Statistical science*, pages 135–140.

Thomas Eiter and Heikki Mannila. 1994. Computing discrete fréchet distance.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade v2: Sparse lexical and expansion model for information retrieval.

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Amin Gheibi, Anil Maheshwari, Jörg-Rüdiger Sack, and Christian Scheffer. 2014. Minimum backward fréchet distance. In *Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 381–388.

GM Harshvardhan, Mahendra Kumar Gourisaria, Manjusha Pandey, and Siddharth Swarup Rautaray. 2020. A comprehensive survey and analysis of generative models in machine learning. *Computer Science Review*, 38:100285.

Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. *arXiv preprint arXiv:1904.02792*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Minghui Jiang, Ying Xu, and Binhai Zhu. 2008. Protein structure–structure alignment with discrete fréchet distance. *Journal of bioinformatics and computational biology*, 6(01):51–64.

Roger W Johnson. 2001. An introduction to the bootstrap. *Teaching statistics*, 23(2):49–54.

Muhammad Zeeshan Khan, Saira Jabeen, Muhammad Usman Ghani Khan, Tanzila Saba, Asim Rehmat, Amjad Rehman, and Usman Tariq. 2020. A realistic image generation of face from text description using the fully trained generative adversarial networks. *IEEE Access*, 9:1250–1260.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. 2023. The role of imagenet classes in fréchet inception distance. In *Proc. ICLR*.

Jimmy Lin and Xueguang Ma. 2021. A few brief notes on deepimpact, coil, and a conceptual framework for information retrieval techniques. *arXiv preprint arXiv:2106.14807*.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021a. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021b. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 163–173.

Joel Mackenzie, Matthias Petri, and Alistair Moffat. 2021. A sensitivity analysis of the msmarco passage collection. *arXiv preprint arXiv:2112.03396*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.

Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019a. From doc2query to docTTTTTquery. *Online preprint*, 6.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019b. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

Artem Obukhov and Mikhail Krasnyanskiy. 2020. Quality assessment method for gan based on modified metrics inception score and fréchet inception distance. In *Software Engineering Perspectives in Intelligent Systems: Proceedings of 4th Computational Methods in Systems and Software 2020, Vol. 1 4*, pages 102–114. Springer.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2010.08191*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. 1995. Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. Gaithersburg, MD: NIST.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2021. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *CoRR*, abs/2112.01488.

Michael Soloveitchik, Tzvi Diskin, Efrat Morin, and Ami Wiesel. 2021. Conditional frechet inception distance. *arXiv preprint arXiv:2103.11521*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *CoRR*, abs/2104.08663.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. 2015. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*.

Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Maarten de Rijke, Haitian Chen, Min Zhang, and Shaoping Ma. 2020. Preference-based evaluation metrics for web image search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 369–378.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval.

Xinyi Yan, Chengxi Luo, Charles LA Clarke, Nick Craswell, Ellen M Voorhees, and Pablo Castells. 2022. Human preferences as dueling bandits. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 567–577.

Li-Chia Yang and Alexander Lerch. 2020. On the evaluation of generative models in music. *Neural Computing and Applications*, 32(9):4773–4784.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022. Scaling autoregressive models for content-rich text-to-image generation.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. Repbert: Contextualized text embeddings for first-stage retrieval. *arXiv preprint arXiv:2006.15498*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.