MS MARCO Web Search: a Large-scale Information-rich Web Dataset with Millions of Real Click Labels

Qi Chen, Microsoft, Beijing, China
Xiubo Geng, Microsoft, Beijing, China
Corby Rosset, Microsoft, Redmond, United States
Carolyn Buractaon, Microsoft, Redmond, United States
Jingwen Lu, Microsoft, Redmond, United States
Tao Shen*, University of Technology Sydney, Sydney, Australia
Kun Zhou, Microsoft, Beijing, China
Chenyan Xiong*, Carnegie Mellon University, Pittsburgh, United States
Yeyun Gong, Microsoft, Beijing, China
Paul Bennett*, Spotify, New York, United States
Nick Craswell, Microsoft, Redmond, United States
Xing Xie, Microsoft, Beijing, China
Fan Yang, Microsoft, Beijing, China
Bryan Tower, Microsoft, Redmond, United States
Nikhil Rao, Microsoft, Mountain View, United States
Anlei Dong†, Microsoft, Mountain View, United States
Wenqi Jiang†, ETH Zürich, Zürich, Switzerland
Zheng Liu†, Microsoft, Beijing, China
Mingqin Li†, Microsoft, Redmond, United States
Chuanjie Liu†, Microsoft, Beijing, China
Zengzhong Li†, Microsoft, Redmond, United States
Rangan Majumder†, Microsoft, Redmond, United States
Jennifer Neville†, Microsoft, Redmond, United States
Andy Oakley†, Microsoft, Redmond, United States
Knut Magne Risvik†, Microsoft, Oslo, Norway
Harsha Vardhan Simhadri†, Microsoft, Bengaluru, India
Manik Varma†, Microsoft, Bengaluru, India
Yujing Wang†, Microsoft, Beijing, China
Linjun Yang†, Microsoft, Redmond, United States
Mao Yang†, Microsoft, Beijing, China
Ce Zhang*†, ETH Zürich, Zürich, Switzerland

ABSTRACT

Recent breakthroughs in large models have highlighted the critical significance of data scale, labels and modals. In this paper, we introduce MS MARCO Web Search, the first large-scale information-rich web dataset, featuring millions of real clicked query-document labels. This dataset closely mimics real-world web document and query distribution, provides rich information for various kinds of downstream tasks and encourages research in various areas, such as generic end-to-end neural indexer models, generic embedding models, and next generation information access system with large language models. MS MARCO Web Search offers a retrieval benchmark with three web retrieval challenge tasks that demands innovations in both machine learning and information retrieval system research domains. As the first dataset that meets large, real and rich data requirements, MS MARCO Web Search paves the way for future advancements in AI and system research. MS MARCO Web Search dataset is available at: https://github.com/microsoft/MS-MARCO-Web-Search.

*The work was done at Microsoft.
†Author names listed alphabetically by surname.

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05
https://doi.org/10.1145/3589335.3648327

CCS CONCEPTS

• **Information systems** → **Relevance assessment**.

KEYWORDS

dataset; information retrieval; web search

ACM Reference Format:

Qi Chen, Xiubo Geng, Corby Rosset, Carolyn Buractaon, Jingwen Lu, Tao Shen, Kun Zhou, Chenyan Xiong*, Yeyun Gong, Paul Bennett*, Nick Craswell, Xing Xie, Fan Yang, Bryan Tower, Nikhil Rao, Anlei Dong, Wenqi Jiang†, Zheng Liu†, Mingqin Li†, Chuanjie Liu†, Zengzhong Li†, Rangan Majumder†, Jennifer Neville†, Andy Oakley†, Knut Magne Risvik†, Harsha Vardhan Simhadri†, Manik Varma†, Yujing Wang†, Linjun Yang†, Mao Yang†, and Ce Zhang*†. 2024. MS MARCO Web Search: a Large-scale Information-rich Web Dataset with Millions of Real Click Labels. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3589335.3648327

1 INTRODUCTION

Recently, the large language model (LLM), a breakthrough in the field of artificial intelligence, has provided a novel way for people to access information through interactive communication. Although it has become an indispensable tool for tasks such as content creation, semantic understanding and conversational AI, it still exhibits certain limitations. One such limitation is the model's tendency to produce hallucinated or fabricated content, as it generates responses based on patterns observed in the training data rather than verifying factual accuracy. Furthermore, it struggles with real-time knowledge updates, as it can only provide information available up until the time of its last training. This makes it less reliable for retrieving the latest, dynamic information. Therefore, integrating an external up-to-date knowledge base with large language models is of paramount importance to enhance their performance and reliability. This combination not only mitigates the limitations of hallucination and knowledge update but also broadens the model's applicability across various domains, making it more versatile and valuable. Consequently, information retrieval systems, like the Bing search engine [31], continue to play a vital role in the new LLM-based information systems, such as Webgpt [33] and new Bing [32].

For modern information retrieval systems, the core is the large semantic understanding model, such as a neural indexer model [50] or dual embedding model [16, 20, 21, 37–39, 44, 45, 53], which can capture users' intents as well as the rich meanings of a document with better tolerance for out of vocabulary words, spelling errors, and synonymous expressions. Training a high-quality large semantic understanding model requires a vast amount of data to achieve sufficient knowledge coverage. The larger the dataset, the better the model is likely to perform, as the model can learn more complex and sophisticated patterns and correlations.

High-quality human-labeled data is as important as data scale. Recent research, such as InstructGPT [35] and LLAMA-2 [49], has demonstrated the crucial role of labeled data for training large foundation models. These models rely on large volumes of training data to learn generalizable features, while human-labeled data enable the model to learn the specific tasks it is designed for. This also applies to large semantic understanding models.

Moreover, information-rich data is also crucial for training large semantic understanding models effectively. The use of multi-modal datasets can help models understand complex relationships between different types of data and transfer knowledge between them. For example, using images and text in a multi-modal data set can help models learn about image concepts and their corresponding text descriptions, providing a more holistic representation of the data.

The emerging large, real and rich data requirements motivate us to create a new MS MARCO Web Search dataset, the first large-scale information-rich web dataset with millions of real clicked query-document labels. MS MARCO Web Search incorporates the largest open web document dataset, ClueWeb22 [36], as our document corpus. ClueWeb22 includes about 10 billion high-quality web pages, sufficiently large to serve as representative web-scale data. It also contains rich information from the web pages, such as visual representation rendered by web browsers, raw HTML structure, clean text, semantic annotations, language and topic tags labeled by industry document understanding systems, etc. MS MARCO Web Search further contains 10 million unique queries from 93 languages with millions of relevant labeled query-document pairs collected from the search log of the Microsoft Bing search engine to serve as the query set. This large collection of multi-lingual information-rich real web documents, queries and labeled query-document pairs enables various kinds of downstream tasks and encourages several new research directions that previous datasets cannot well support, for example, generic end-to-end neural indexer models, generic embedding models, and next generation information access system with large language models, etc. As the first large, real and rich web dataset, MS MARCO Web Search will serve as a critical data foundation for future AI and systems research.

MS MARCO Web Search offers a retrieval benchmark which implements several state-of-the-art embedding models, retrieval algorithms, and retrieval systems originally developed on existing datasets. We compare the quality of their results and system performance on our new MS MARCO Web Search dataset as the benchmark baselines for web scale information retrieval. The experiment results demonstrate that embedding models, retrieval algorithms, and retrieval systems are all critical components in web information retrieval. And interestingly, simply improving only one component may bring negative impacts to the end-to-end retrieval result quality and system performance. We hope that this retrieval benchmark can facilitate future innovations in data-centric techniques, embedding models, retrieval algorithms, and retrieval systems to maximize end-to-end performance.

2 BACKGROUND AND RELATED WORK

2.1 Web Scale Information Retrieval

In traditional information retrieval, user queries and documents are represented as a list of keywords, and the retrieval is done based on keyword matching. However, simple keyword matching faces many challenges. First, it cannot clearly understand users' intents. In particular, it cannot estimate users' positive and negative sentiment and may return opposite results by mistake. Second, it cannot combine synonymous expressions, reducing the diversity of results [18]. Third, it cannot handle spelling errors and will return irrelevant results. Therefore, query alteration is employed to address

Table 1: Comparison of MS MARCO Web Search (with ClueWeb22) and existing datasets

Dataset	#Documents	#Queries	Web	Multi-lingual docs	Rich-info docs	Multi-lingual queries
Robust04	528K	250	-	-	-	-
ClueWeb09	1B	-	✓	✓	-	-
ClueWeb12	733M	-	✓	-	-	-
GOV2	25M	50	-	-	-	-
Common Crawl (One Dump)	3.1B	-	✓	✓	-	-
Natural Questions	28M	320K	-	-	-	-
MS MARCO	3.2M	100K	-	-	-	-
MS MARCO Ranking v2	11M	1M	-	-	-	-
ORCAS	3.2M	10M	-	-	-	-
CLIR	23.9M	2.8M	-	✓	-	-
MS MARCO Web Search (w. ClueWeb22)	**10B**	**10M**	✓	✓	✓	✓

the above challenges. Unfortunately, it is difficult to cover all kinds of query alterations, especially those newly-appeared alterations.

With the great success of deep learning in natural language processing, both queries and documents can be more meaningfully represented as semantic embedding vectors. Since embedding-based retrieval solves the above three challenges, it has been widely used in modern information systems to facilitate new state-of-the-art retrieval quality and performance. Numerous prior studies have concentrated on deep embedding models, from DSSM [21], CDSSM [45], LSTM-RNN [37], and ARC-I [20] to transformer-based embedding models [10, 16, 38, 39, 44, 52, 53]. They have shown impressive gains with brute-force nearest neighbor embedding search on some small datasets as compared with traditional keyword matching.

Due to the extremely high computational cost and query latency of brute-force vector search, there are many research approaches focusing on large-scale approximate vector nearest neighbor search (ANN) algorithms and systems design [5–7, 11, 19, 23–25, 25, 40, 47]. They can be divided into partition-based and graph-based solutions. Partition-based solutions, such as SPANN [11], divide the whole vector space into a large number of clusters and only do fine-grained search on a small number of the closest clusters to a query in online search. Graph-based solutions, such as DiskANN [47], construct a neighbor graph for the whole dataset and do the best-first traversal from some fixed starting points when a query comes in. Both of these approaches work well on some uniform-distributed datasets.

Unfortunately, when applying embedding-based retrieval in the web scenario, several new challenges emerge. First, web scale data volumes require large models, high embedding dimensions, and a large-scale labeled training dataset to guarantee sufficient knowledge coverage. Second, performance gains of state-of-the-art embedding models verified on small datasets cannot be directly transferred to a web scale dataset (see section 4.4). Third, embedding models need to co-work with ANN systems in order to serve large scale data volumes efficiently. However, different training data distributions may affect the accuracy and system performance of an ANN algorithm, which will greatly reduce the result accuracy as compared to embedding models with brute-force search. Distill-VQ [51] has verified that CoCondenser [17] embedding model with Faiss-IVFPQ ANN index achieves different result accuracy on MSMarco [34] and NQ [27] datasets. Moreover, even the same training data distribution will also result in different embedding vector distributions, which will lead to different ranking trends of the embedding models in brute-force search (KNN) and approximate nearest neighbor search (ANN) (see section 4.6).

2.2 Existing Datasets

To encourage innovation in the information retrieval area, the community has collected several datasets for public benchmarking (summarized in Table 1).

There are many public web datasets for traditional information retrieval tasks, such as Robust04 [3], ClueWeb09 [13], ClueWeb12 [9], GOV2 [12], ClueWeb22 [36] and Common Crawl [2]. Unfortunately, these datasets have, at most, hundreds of labeled queries, far from enough to learn a good deep learning enhanced retrieval model.

Recently, several new datasets have been published for research on deep learning enhanced retrieval [27, 34, 42]. MS MARCO [34] is one of the most popular datasets for embedding model investigation. It provides 100K questions collected from Bing's search questions paired with human generated answers contextualized within web documents. MS MARCO Ranking v2 [46] expands the size of the document and question sets to 11 million and 1 million, respectively. ORCAS [14] provides 10 million unique queries and 18 million clicked query-document pairs for MS MARCO documents. Natural Questions [27], a sub-million-scale question answering dataset collected from Google's search queries with human annotated answers in Wikipedia articles, was repurposed for embedding-based retrieval by extracting passages from Wikipedia as candidate answers [26]. CLIR [42], a million-scale cross-language information retrieval dataset collected from Wikipedia, has been used to train cross-lingual embedding models [54]. However, none of these datasets meet the emerging large, real and rich requirements. These datasets focus on English-only question answering tasks. None of them has the desired web-scale data with highly-skewed multilingual queries which can be short, ambiguous and often not formulated as natural language questions. Further, they only provide the raw text of queries and answers, which limits the potential of future cross-modal knowledge transfer research. Finally, they only focus on evaluating the quality of embedding models using brute-force search, which cannot reflect end-to-end retrieval challenges.

ANN benchmark [4] and Billion-scale ANN benchmark [1] provide multiple high-dimensional vector datasets to evaluate the result accuracy and system performance for embedding-based retrieval algorithms. Unfortunately, they cannot measure model quality and thus cannot reflect the end-to-end retrieval performance.

Therefore, a large-scale information-rich web dataset with real document and query distribution that can reflect real-world challenges is still lacking.

3 MS MARCO WEB SEARCH DATASET

In this paper, we present MS MARCO Web Search, a large-scale dataset for research on web information retrieval. MS MARCO Web Search dataset consists of a high quality set of web pages that mirrors the highly-skewed web document distribution, a query set that reflects the real web query distribution, and a large-scale query-document label set for embedding model training and evaluation.

3.1 Document Preparation

We use ClueWeb22 [9] as our document set since it is the largest and newest open web document dataset for our purpose. It meets the requirements of large scale, high quality and realistic document distributions crawled and processed by a commercial web search engine with rich information. Compare to Common Crawl [2] which only crawls 35 million registered domains and covers 40+ languages, ClueWeb22 closely mimics the realistic crawl selection of a commercial search engine with 207 languages. It has 10 billion high-quality web pages with rich affiliated information, such as url, language tag, topic tag, title and clean text, etc. Figure 1 (d) gives an example of the data structures provided by ClueWeb22.

To make training cost-effective for both academia and industry, we provide a 100 million and a 10 billion document set. The 100 million document set is a random subset of the 10 billion document set. In order to evaluate model generalization ability in a small-scale dataset, two 100 million non-overlapping document sets are provided, one for training and the other for testing.

3.2 Query Selection and Labeling

To generate large scale high quality queries and query-document relevance labels, we sample query-document clicks from one year of Bing search engine's logs. The initial query set gets filtered to remove queries that are rarely triggered, contain personally identifiable information, offensive content, adult content and those having no click connection to the ClueWeb22 document set. The resulting set includes queries triggered by many users, which reflects the real query distribution of a commercial web search engine.

The queries are split into train and test sets based on time, which is similar to real-world web scenarios training an embedding model using past data and serving future incoming web pages and queries. We sample around 10 million query-document pairs from the train set and 10 thousand pairs from the test set. The documents in the query-document train and test sets are then merged into the 100 million train document set and test document set respectively. To enable quality verification of the model during training, we split a dev query-document set from the train query-document set. Since the train and dev sets share the same document set, the dev set can be used to quickly verify the training correctness and model quality during training. For the 10B dataset, we use the same train, dev, and test queries but sample more query-document pairs.

3.3 Dataset Analysis

We have constructed two scales of the datasets: Set-100M and Set-10B. Table 2 gives the detailed statistics of the datasets. The example files of MS MARCO Web Search Set-100M are shown in figure 1.

3.3.1 Language Distribution Analysis. MS MARCO Web Search is a multi-lingual dataset with its queries and documents both from a commercial web search engine. We analyze the 20 most popular languages among 93 and 207 languages in both queries and documents in the 100M dataset respectively; the 10B dataset has a similar distribution. Figure 2 demonstrates the document language distribution in the train and test document sets are aligned with the original ClueWeb22 document distribution. Figure 3 summarizes the query language distribution in the train, dev, and test query sets. From the distribution, we can see that the language distribution of the queries in the web scenario is high-skewed which may lead to model bias. It encourages research on data-centric techniques for training data optimization.

3.3.2 Data Skew Analysis. We analyze the query-document label distribution in the training data. Figure 4(a) shows documents and the number of relevant queries associated with them. From the figure, we can see that there are only a few documents with multiple labels: only 7.77% of the documents have relevant labeled queries and 0.46% of documents have more than one labeled relevant query. Figure 4(b) summarizes the queries and their relevant documents. From the figure, we can see that only 1.4% of queries have multiple relevant documents. This highly skewed nature of the dataset is consistent with what is observed while training models for web-scale information retrieval. Our intention is to keep this skew to make models trained on this dataset applicable to real-world scenarios.

3.3.3 Test-Train Overlap Analysis. As introduced in [29], there exists large test-train overlap in some popular open-domain QA datasets, which cause many popular open-domain models to simply memorize the queries seen at the training stage. Subsequently, they perform worse on novel queries. The work [55] observes this phenomenon in the MSMARCO dataset. To better evaluate model generalizability, we minimize the overlap between the train and test sets by splitting the query-document pairs into train and test sets by time. This means the test query-document pairs have no time overlap with the train query-document pairs, which introduces a large portion of novel queries. This can be verified in the table 3. We summarize the test query-document pairs into four categories:

- **Q∈Train, D∈Train**: Both query and document have appeared in the train set,
- **Q∉Train, D∈Train**: Query has not been seen in the train set, but the relevant document has been seen in the train set,
- **Q∈Train, D∉Train**: Query has been seen in the train set, but the document is a new web page that has not been seen in the train set,
- **Q∉Train, D∉Train**: Both query and document are novel content which have never been seen in the train set.

We can see from the table 3 that 82% of query-document pairs are novel content in the test set which have not been seen in the

Qid	Query	Language
1	40 inch tv price in india	en-IN, en-US, en-GB, en-IN, en-US
2	++i and i++ difference	en-MY, en-SG, en-US, en-WW, fr-FR, en-IN
...	...	...
10	设计一个基于物联网技术的智能家居系统，画出结构框图，并针对其中一个功能画出流程图。	zh-CN

(a) queries_[test/train/dev].tsv

Qid	_	DocId	Score
7952	0	155074134	1
1452	0	102977197	1
...	...	...	...
8989	0	157326226	1

(b) qrels_[test/train/dev].tsv

URLHash	DocId
001FB55DA2FAE40D352C26DEE71C7204	1
BA7EE9F93BCA48504804E0688F72B525	2
...	...
CEAF10C6CA43A70D8ED1C183E53EA5A3	10

(c) doc_hash_mapping.tsv

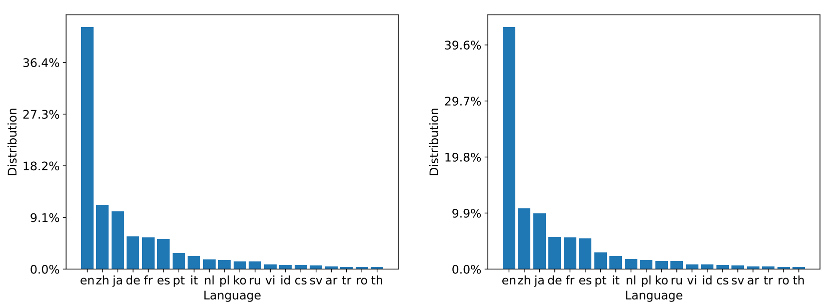
URL	URLHash	Language	Topic	Title	Primary
https://bvca.fanya.chaoxing.com/portal	0000000D67F0691EF63990F85F8794CB	zh_chs	"china", "education“, ...	北京农业职业学院网络教学平台	北京农业职业学院网络教学平台\n尔雅网络通识课\n0学分 热度\n...
https://www.berliner-zeitung.de/news/was-das-streaming-und-fernsehjahr-2022-bringt-li.203314	0000007BBDFF042D81647CBC9A41C16C	de	"amazon prime", "television“, ...	Was das Streaming- und Fernsehjahr 2022 bringt	Was das Streaming- und Fernsehjahr 2022 bringt\nNetflix, Amazon Prime, TV und Co\n...
...	...	...	...	...	...
https://www.casasbahia.com.br/lavadora-lg-vc2/b	000000D9C6F4F89D6C53AE593B71BDE2	pt	"lg mobile", "smartphones“, ...	Você buscou por lavadora lg vc2	Lavadora lg vc2 | Casas Bahia\nVocê buscou por lavadora lg vc2\n...

(d) documents_[train/test].tsv from ClueWeb22

Figure 1: The example files in the MS MARCO Web Search dataset

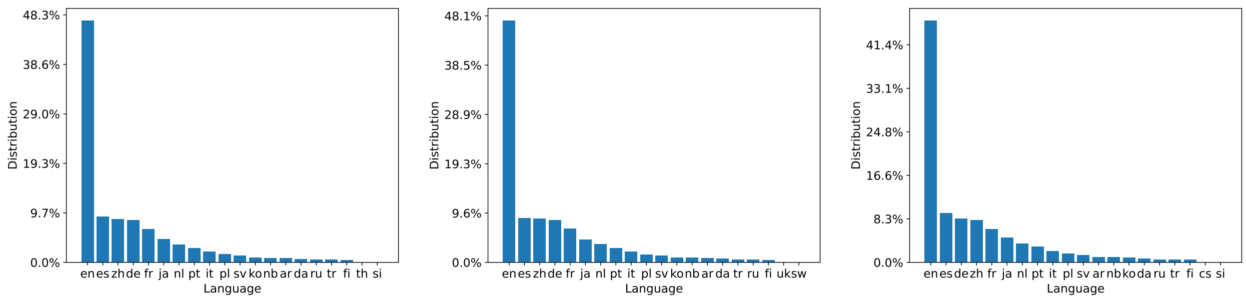
Table 2: 100M and 10B dataset statistics

Scale	Train			Dev			Test		
	Document	Query	Q-D	Document	Query	Q-D	Document	Query	Q-D
Set-100M	109969872	9206475	9346695	109969872	9253	9402	100924960	9374	9374
Set-10B	10B	9206475	62302553	10B	9253	63314	10B	9374	40511



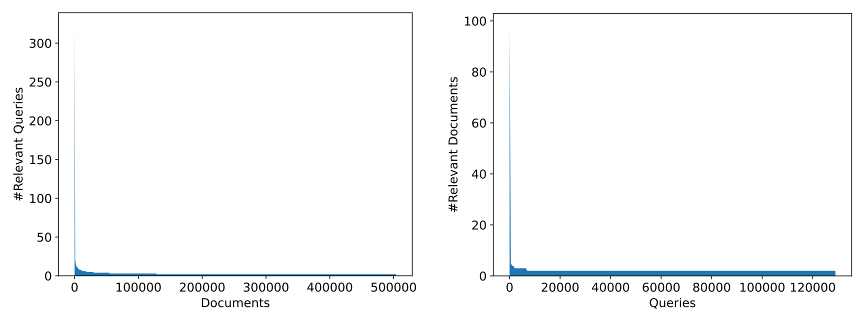
(a) Train (b) Test

Figure 2: The distribution of 20 most popular languages in the train and test document sets.



(a) Train (b) Dev (c) Test

Figure 3: The distribution of 20 most popular languages in the train, dev and test query sets.



(a) (b)

Figure 4: The distribution of documents and queries labels

Table 3: Test query-document pair distribution

Total	Q∈Train, C∈Train	Q∉Train, C∈Train	Q∈Train, C∉Train	Q∉Train, C∉Train
9374	82	1468	178	7646

train set. Therefore, MS MARCO Web Search dataset is capable of offering effective assessments of models based on memory capacity and generalizability by dividing the test set into four categories for a more detailed comparison.

3.4 New Challenges Raised by MS MARCO Web Search

Based on the MS MARCO Web Search datasets, we raise three challenge tasks in large embedding model and retrieval system design.

3.4.1 Large-scale Embedding Model Challenge. As introduced before, the large-scale web data volume requires large embedding models to guarantee sufficient knowledge coverage. It requires balancing the following two goals: good model generalization ability and efficient train/inference speed.

3.4.2 Embedding Retrieval Algorithm Challenge. Embedding models need to co-work with the embedding retrieval system to serve a web scale dataset. In this challenge, we take the embedding vectors generated by our best baseline model as the ANN vector set. The goal of this challenge is to call for ANN algorithm innovations to minimize the accuracy gap between approximate search and brute-force search while still preserving good system performance.

3.4.3 End-to-end Retrieval System Challenge. In the web scenario, the result quality and system performance of the end-to-end retrieval system are the most important metrics in comparing different solutions. This challenge task encourages any kind of solutions, including an embedding model plus ANN system [30], inverted index solution [8, 15, 57], hybrid solution [18, 28, 43], neural indexer [48, 50], and large language model [49] etc.

4 BENCHMARK RESULTS

In this section, we provide initial benchmark results for some state-of-the-art embedding models, ANN algorithms, and popular retrieval systems on the MS MARCO Web Search 100M dataset as baselines. For the 10B dataset, we leave it for open exploration.

Table 4: Result quality of baseline models

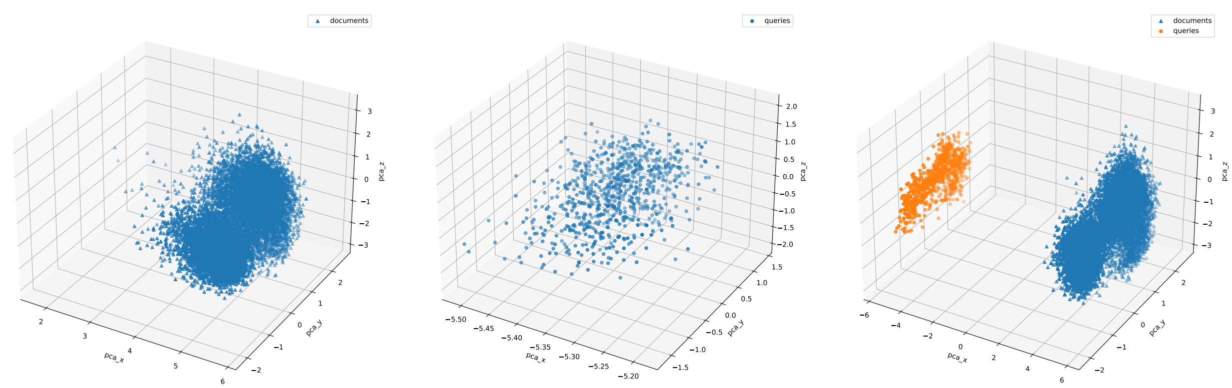
Baselines	MS MARCO Web Search						NQ		MS MARCO Passage	
	MRR@10	recall@1	recall@5	recall@10	recall@20	recall@100	recall@20	recall@100	recall@50	recall@1K
DPR	0.542	45.12%	66.04%	72.10%	76.80%	87.54%	78.4%	85.4%	-	-
ANCE	0.633	54.18%	75.53%	80.53%	84.17%	91.17%	81.9%	87.5%	81.1%	95.9%
SimANS	**0.649**	**55.86%**	**76.84%**	**81.78%**	**85.23%**	**91.98%**	**86.2%**	**90.3%**	**88.7%**	**98.7%**

Table 5: Performance of ANN baselines

Baselines	ANN recall@1	ANN recall@10	ANN recall@100	QPS	P50 latency	P90 latency	P99 latency
SPANN	87.97%	80.55%	69.84%	625	10.411 ms	10.873 ms	11.334 ms
DiskANN	91.46%	87.07%	69.73%	2691	21.968 ms	37.841	69.462 ms

Table 6: Result quality and system performance of baseline systems

Baselines	MRR@10	recall@1	recall@5	recall@10	recall@20	recall@100	QPS	P50 latency	P90 latency	P99 latency
BM25	0.296	22.30%	39.04%	46.00%	52.42%	63.87%	149	312.025 ms	1065.141 ms	3745.546 ms
DPR	0.467	39.21%	56.66%	61.27%	64.69%	70.28%				
ANCE	0.580	49.87%	68.59%	72.94%	75.86%	80.18%	625	21.924 ms	23.017 ms	34.217 ms
SimANS	0.585	50.63%	68.79%	73.14%	75.85%	79.82%				



(a) Documents (b) Queries (c) Documents & Queries

Figure 5: The large gap between embedding vector distribution of documents and queries in SimANS.

4.1 Environment Setup

We use the Azure Standard_ND96asr_v4 Virtual Machine for model training and performance testing. It contains 96 vCPU cores, 900 GB memory, 8 A100 40GB GPUs with NVLink 3.0.

4.2 Baseline Methods

4.2.1 State-of-the-art Embedding Models. We take the following three state-of-the-art multi-lingual models as the initial baseline models:

- DPR [26] is based on a BERT pre-trained model and a dual-encoder architecture, whose embedding is optimized for maximizing the inner product score of a query and its relevant passage. The negative training examples are selected by BM25 retrieved documents.
- ANCE [53] improves embedding-based retrieval performance by selecting hard negative training examples from the entire document set using an asynchronously updated approximate nearest neighbor index.
- SimANS [56] avoids over indexing on false negatives by selecting ambiguous samples rather than the hardest ones.

We aspire for the MS MARCO Web Search dataset to establish itself as a new standard benchmark for retrieval, enticing more baseline models to assess and experiment with it.

4.2.2 State-of-the-art ANN Algorithms. For embedding retrieval algorithms, we choose the state-of-the-art disk-based ANN algorithms DiskANN [22] and SPANN [11] as the baselines.

4.2.3 End-to-end Retrieval Systems. BM25 [41] is the most widely used ranking function in the web information retrieval area to estimate the relevance score of a document given a query. It ranks a set of documents based on a probabilistic retrieval framework and has been integrated into the Elasticsearch system to serve all kinds of search scenarios.

4.3 Evaluation Metrics

We evaluate all the baselines on both result quality and system performance aspects. For result quality, we take Mean Reciprocal Rank (MRR) and recall as the evaluation metrics:

- MRR: the average of the multiplicative inverse of the rank of the first correct result, which is widely used for evaluating the model quality.
- Recall: the average percentage of ground truth items recalled during the search. For the embedding model challenge and the end-to-end retrieval system challenge, we use our test query-document labels as the ground truth. For the embedding retrieval algorithm challenge, we use the brute-force vector search results as the ground truth (ANN recall) to evaluate the ANN algorithm performance.

For system performance, we evaluate the following metrics under limited resource cost to align with industry scenarios:

- Throughput: All queries are provided at once, and we measure the wall clock time between the ingestion of the vectors and when all the results are output using all the threads in a machine. Then the throughput is calculated as the processed queries per second (QPS).
- Latency: we measure the 50, 90 and 99 percentile query latency at certain QPS.

4.4 Evaluation of Embedding Models

In this experiment, we measure the MRR and recall of all the baseline embedding models. From the result, we can see that SimANS with the ambiguous samples as negative training examples performs the best on MS MARCO Web Search 100M dataset. The ranking of the baseline models is aligned with the model evolution trend in the literature. Nonetheless, when compared with the evaluation results in Natural Question (NQ) [27] and MS MARCO Passage Ranking [34], the gap in performance between ANCE and SimANS in MS MARCO Web Search becomes less significant.

We also evaluate the system performance for the three baseline embedding models. Since they use the same model architecture and the same number of parameters, their serving time cost is similar. At the peak of 698 QPS, the latency percentiles of 50, 90, and 99 are 9.896 ms, 10.018 ms, and 11.430 ms, respectively.

4.5 Evaluation of ANN Algorithms

In this experiment, we evaluate the ANN performance with vectors generated by the best baseline model. We build both DiskANN and SPANN indices and evaluate both their serving performance and result quality. Here we only focus on evaluating the gap between ANN and KNN. Therefore, we use the brute-force search results as the ground truth to measure recall. Table 5 summarizes the recall and system performance of the two baselines. From the results, we can see that it is difficult to achieve high recall when the number of return results K is large. One of the reasons is that the distributions of queries and documents are highly-skewed and far away from each other (see figure 5). We also observe this phenomenon in DPR and ANCE embeddings.

4.6 Evaluation of End-to-end Performance

In this section, we evaluate the end-to-end performance of the three baseline embedding models plus SPANN index and the widely-used Elasticsearch BM25 solution. Table 6 demonstrate the result quality and system performance of all these baseline systems, respectively. Compared with table 4, we can see that after using the ANN index, the final result quality drops a lot. For example, the metric recall@100 drops more than 10 points for all baseline models. There exists large quality gaps between the ANN and KNN results (see table 5). Moreover, we notice that using the ANN index will change the model ranking trend. SimANS achieves the best results for all the result quality metrics with brute-force search. However, when using the SPANN index, it performs worse than ANCE in recall@20 and recall@100. We further analyze the phenomenon in detail and find that SimANS has a larger gap between the average distance of query to the top100 documents relative to the average distance of document to the top100 documents than ANCE. The gap in SimANS and ANCE are 103.35 and 73.29, respectively. This will cause inaccurate distance bound estimation for a query to the neighbors of a document. As a result, ANN cannot perform well because it relies on distance estimated according to the triangle inequality. Both result quality and system performance results of the end-to-end evaluation call for more innovations on the end-to-end retrieval system design.

5 POTENTIAL BIASES AND LIMITATIONS

As discussed in section 3.3.1, The language distribution of documents and queries in the web scenario is high-skewed. This will lead to language bias on data and models. ClueWeb22 [9] demonstrates that there also exists topic distribution skew in the web scenario. Therefore, domain bias also may happen in data and models.

To protect user privacy and content health, we remove queries that are rarely triggered (triggered by less than K users, where K is a high value), contain personally identifiable information, offensive content, adult content and queries that have no click connection to the ClueWeb22 document set. As a result, the query distribution is slightly different from the real web query distribution.

6 FUTURE WORK AND CONCLUSIONS

MS MARCO Web Search is the first web dataset that effectively meets the criteria of being large, real, and rich in terms of data quality. It is composed of large-scale web pages and query-document labels sourced from a commercial search engine, retaining rich information about the web pages that is widely employed in industry. The retrieval benchmark offered by MS MARCO Web Search comprises three challenging tasks that require innovation in both the areas of machine learning and information retrieval system research. We hope MS MARCO Web Search can serve as a benchmark for modern web-scale information retrieval, facilitating future research and innovation in diverse directions.

REFERENCES

[1] [n. d.]. Billion-scale ANNS Benchmarks. https://big-ann-benchmarks.com/.

[2] [n. d.]. Common Crawl.

[3] [n. d.]. Robust04. https://trec.nist.gov/data/robust/04.guidelines.html.

[4] Martin Aumüller, Erik Bernhardsson, and Alexander Faithfull. 2017. ANN-benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. In *International conference on similarity search and applications*. Springer, 34–49.

[5] Artem Babenko and Victor Lempitsky. 2014. The inverted multi-index. *IEEE transactions on pattern analysis and machine intelligence* 37, 6 (2014), 1247–1260.

[6] Artem Babenko and Victor Lempitsky. 2016. Efficient indexing of billion-scale datasets of deep descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2055–2063.

[7] Dmitry Baranchuk, Artem Babenko, and Yury Malkov. 2018. Revisiting the inverted indices for billion-scale approximate nearest neighbors. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 202–216.

[8] Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers. *Advances in Neural Information Processing Systems* 35 (2022), 31668–31683.

[9] Jamie Callan. 2012. The lemur project and its clueweb12 dataset. In *Invited talk at the SIGIR 2012 Workshop on Open-Source Information Retrieval*.

[10] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216* (2024).

[11] Qi Chen, Bing Zhao, Haidong Wang, Mingqin Li, Chuanjie Liu, Zengzhong Li, Mao Yang, and Jingdong Wang. 2021. SPANN: Highly-efficient Billion-scale Approximate Nearest Neighborhood Search. *Advances in Neural Information Processing Systems* 34 (2021), 5199–5212.

[12] Charles Clarke, Nick Craswell, and Ian Soboroff. 2004. Overview of the TREC 2004 Terabyte Track. In *TREC*.

[13] Charles LA Clarke, Nick Craswell, and Ian Soboroff. 2009. Overview of the TREC 2009 Web Track.. In *Trec*, Vol. 9. 20–29.

[14] Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. 2020. ORCAS: 20 million clicked query-document pairs for analyzing search. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2983–2989.

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

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[17] Luyu Gao and Jamie Callan. 2022. Unsupervised Corpus Aware Language Model Pre-training for Dense Passage Retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2843–2853.

[18] Jiafeng Guo, Yinqiong Cai, Yixing Fan, Fei Sun, Ruqing Zhang, and Xueqi Cheng. 2022. Semantic models for the first-stage retrieval: A comprehensive review. *ACM Transactions on Information Systems (TOIS)* 40, 4 (2022), 1–42.

[19] Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. 2020. Accelerating Large-Scale Inference with Anisotropic Vector Quantization. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*. 3887–3896.

[20] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. *Advances in neural information processing systems* 27 (2014).

[21] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2333–2338.

[22] Suhas Jayaram Subramanya, Fnu Devvrit, Harsha Vardhan Simhadri, Ravishankar Krishnawamy, and Rohan Kadekodi. 2019. Diskann: Fast accurate billion-point nearest neighbor search on a single node. *Advances in Neural Information Processing Systems* 32 (2019).

[23] Hervé Jégou, Romain Tavenard, Matthijs Douze, and Laurent Amsaleg. 2011. Searching in one billion vectors: re-rank with source coding. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 861–864.

[24] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* (2019).

[25] Yannis Kalantidis and Yannis Avrithis. 2014. Locally optimized product quantization for approximate nearest neighbor search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2321–2328.

[26] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906* (2020).

[27] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466.

[28] Carlos Lassance and Stéphane Clinchant. 2023. Naver Labs Europe (SPLADE)@ TREC Deep Learning 2022. *arXiv preprint arXiv:2302.12574* (2023).

[29] Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 1000–1008.

[30] Wenhao Lu, Jian Jiao, and Ruofei Zhang. 2020. Twinbert: Distilling knowledge to twin-structured compressed BERT models for large-scale retrieval. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2645–2652.

[31] microsoft. [n. d.]. Bing search. https://www.bing.com/.

[32] microsoft. [n. d.]. New Bing. https://www.bing.com/new.

[33] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332* (2021).

[34] 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. In *CoCo@ NIPS*.

[35] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[36] Arnold Overwijk, Chenyan Xiong, and Jamie Callan. 2022. ClueWeb22: 10 billion web documents with rich information. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3360–3362.

[37] Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. 2016. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, 4 (2016), 694–707.

[38] Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2019. Understanding the Behaviors of BERT in Ranking. *arXiv preprint arXiv:1904.07531* (2019).

[39] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[40] Jie Ren, Minjia Zhang, and Dong Li. 2020. HM-ANN: Efficient Billion-Point Nearest Neighbor Search on Heterogeneous Memory. In *In Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vol. 33.

[41] Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*. Springer, 232–241.

[42] Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. 2018. Cross-lingual learning-to-rank with shared representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 458–463.

[43] Minjoon Seo, Jinhyuk Lee, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi, and Hannaneh Hajishirzi. 2019. Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4430–4441.

[44] Xuan Shan, Chuanjie Liu, Yiqian Xia, Qi Chen, Yusi Zhang, Kaize Ding, Yaobo Liang, Angen Luo, and Yuxiang Luo. 2021. GLOW: Global Weighted Self-Attention Network for Web Search. In *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 519–528.

[45] Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd international conference on world wide web*. 373–374.

[46] Ian Soboroff. 2021. Overview of TREC 2021. In *30th Text REtrieval Conference. Gaithersburg, Maryland*.

[47] Suhas Jayaram Subramanya, Rohan Kadekodi, Ravishankar Krishaswamy, and Harsha Vardhan Simhadri. 2019. Diskann: Fast accurate billion-point nearest neighbor search on a single node. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 13766–13776.

[48] Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, et al. 2022. Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems* 35 (2022), 21831–21843.

[49] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[50] Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, et al. 2022. A neural corpus indexer for document retrieval. *Advances in Neural Information Processing Systems* 35 (2022), 25600–25614.

[51] Shitao Xiao, Zheng Liu, Weihao Han, Jianjin Zhang, Defu Lian, Yeyun Gong, Qi Chen, Fan Yang, Hao Sun, Yingxia Shao, et al. 2022. Distill-vq: Learning retrieval oriented vector quantization by distilling knowledge from dense embeddings. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1513–1523.

[52] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597* (2023).

[53] 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. *arXiv preprint arXiv:2007.00808* (2020).

[54] Linlong Xu, Baosong Yang, Xiaoyu Lv, Tianchi Bi, Dayiheng Liu, and Haibo Zhang. 2021. Leveraging Advantages of Interactive and Non-Interactive Models for Vector-Based Cross-Lingual Information Retrieval. *arXiv preprint arXiv:2111.01992* (2021).

[55] Jingtao Zhan, Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2022. Evaluating Interpolation and Extrapolation Performance of Neural Retrieval Models. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2486–2496.

[56] Kun Zhou, Yeyun Gong, Xiao Liu, Wayne Xin Zhao, Yelong Shen, Anlei Dong, Jingwen Lu, Rangan Majumder, Ji-Rong Wen, and Nan Duan. 2022. SimANS: Simple Ambiguous Negatives Sampling for Dense Text Retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 548–559.

[57] Shengyao Zhuang, Hang Li, and G. Zuccon. 2021. Deep Query Likelihood Model for Information Retrieval. In *ECIR*.