INFINI Labs
极限科技，让搜索更简单

DataFun.

给 ES 插上向量检索的翅膀

利用向量技术提升搜索与检索的能力

张 磊

极限科技 Easysearch 研发工程师

INFINI Labs
极限科技，让搜索更简单

摘要和大纲

- 向量数据库介绍
- 当前向量技术的介绍
- Elasticsearch 介绍
- ES 与向量技术融合后的场景介绍
- 如何利用向量提升 ES搜索能力

INFINI Labs
极限科技，让搜索更简单

向量数据库介绍

向量数据库的主要特点是能够高效地存储和查询大规模的向量数据。它通常采用基于向量相似度的查询方式，即根据向量之间的相似度来检索数据。这种查询方式可以用于各种应用场景，例如图像搜索、音乐推荐、文本分类等。

传统的数据库主要关注存储和查询标量数据，例如文本、数字或日期等。但是，在许多应用领域，如机器学习、计算机视觉、自然语言处理和推荐系统等，需要处理和分析的是向量数据。向量数据库就是为了满足这些需求而设计的。

INFINI Labs
极限科技，让搜索更简单

向量数据库的功能和主要特点

向量数据库的应用范围广泛，包括图像搜索、推荐系统、人脸识别、嵌入式设备、工业物联网等领域。它们为处理和分析向量数据提供了高效、便捷和可 扩展的解决方案。

存储向量数据

向量数据库能够高效地存储大规模的向量数据集，保持向量的完整性和准确性。

高维向量索引

向量数据库使用特殊的索引结构，如树状结构（如KD-Tree、M-Tree）或哈希结构（如LSH、HNSW），以支持高效的向量相似性搜索。

向量相似性搜索

向量数据库能够根据给定的查询向量，快速找到与之相似的向量，以支持近似搜索和相关性分析。

向量聚类和分类

向量数据库提供了聚类和分类算法，可以将向量数据分组成具有相似特征的集合，以支持数据分析和挖掘。

扩展性和性能

向量数据库通常具有良好的可扩展性和高性能，能够处理大规模的向量数据集，并提供快速的查询响应时间。

INFINI Labs
极限科技，让搜索更简单

向量数据库产品

Faiss

Faiss是由Facebook AI研究团队开发的库，用于高效相似度搜索和聚类高维向量。Faiss支持基于CPU和GPU的操作，并且其设计允许在大型数据集中进行最近邻搜索。不过，它并不提供持久化存储和分布式计算功能。

milvus

Milvus 是一款云原生向量数据库，它具备高可用、高性能、易拓展的特点，用于海量向量数据的实时召回。

Pinecone

Pinecone 可以轻松地为高性能人工智能应用提供长期记忆。它是一个托管的云原生矢量数据库，具有简单的 API，并且没有基础设施麻烦。Pinecone 在数十亿个向量的规模上以低延迟提供新鲜的、经过过滤的查询结果。

向量检索技术介绍 / 单词向量化

技术	描述	优点
Word2Vec	word2vec不是一个单一的算法，而是一系列模型架构和优化，可用于从大型数据集中学习单词嵌入。通过word2vec学习的嵌入已被证明在各种下游自然语言处理任务中是成功的。	可以捕捉词之间的语义语法关系，向量紧凑，维度少，无监督，训练速度快。
GloVe	一种基于全局词频统计的词嵌入模型。利用整个语料库中单词的共现统计信息，构建了一个词-词共现矩阵。然后通过奇异值分解(SVD)等方法，将词-词共现矩阵转换为低维的单词向量表示。	可以捕获更长范围的词语依赖关系，可以获取全局语义信息。
FastText	FastText 是Facebook研究团队创建的一个库，用于高效计算词表示和执行文本分类。	能够处理单词多义性，歧义性，尤其在大规模多类别的文本分类任务中，具有明显的优势。FastText还可以处理词汇表外的单词(OOV)。

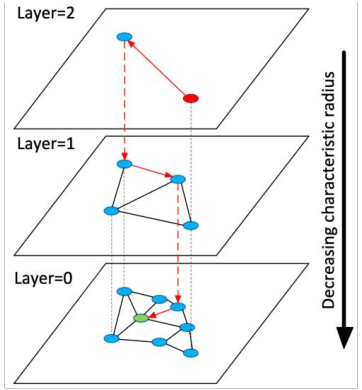
INFINI Labs
极限科技，让搜索更简单

向量检索技术介绍 / 句子向量化

技术	描述
Doc2Vec	Doc2Vec 是一种自然语言处理模型，它用于生成文档或段落的向量表示，是Word2Vec 的扩展。
SentenceBERT	（SBERT）是一个基于BERT模型的变体，能直接为整个句子生成向量表示。相比原始的BERT，SBERT使用孪生网络结构进行训练，这种结构使得模型可以计算句子间的相似度来优化参数
InferSent	InferSent是一个由Facebook AI研究团队开发的句子编码模型，用于生成代表句子含义的向量表示。它是通过监督学习训练的，特别是通过一个叫做自然语言推理（NLI）的任务。
Universal Sentence Encoder	(USE) 是由Google AI开发的一种模型，其目的是为句子提供高质量的语义表示（向量）。这些句子级别的向量表示可以被用于各种自然语言处理任务，如句子相似性、文本聚类或情感分析等。

INFINI Labs
极限科技，让搜索更简单

向量检索技术介绍 / 向量索引结构算法

Hierarchical Navigable Small World **Lucene 9.0 引入 HNSW 来存储向量** **详见** LUCENE-9004 **ES 也是在这个基础上推出了 kNN search**	基于图的数据结构，它将节点划分成不同层级，贪婪地遍历来自上层的元素，直到达到局部最小值，然后切换到下一层，以上一层中的局部最小值作为新元素重新开始遍历，直到遍历完最低一层。 与NSW相比，HNSW的改进方法： 使用分层的结构 使用了一种启发式方法选择某节点的邻居。	

INFINI Labs
极限科技，让搜索更简单

ES 与向量技术融合后的场景

Elasticsearch

是一个免费且开放的分布式搜索和分析引擎，适用于包括文本、数字、地理空 间、结构化和非结构化数据等在内的所有 类型的数据。Elasticsearch 在 Apache Lucene 的基础上开发而成，由 Elasticsearch N.V.（即现在的 Elastic）于 2010 年首次发布。Elasticsearch 以其简单的 REST 风格 API、分布式特性、速度和可 扩展性而闻名.

结合ES 与向量搜索技术可以实现更复杂和强大的文本搜索功能。

8.0 开始引入 kNN search

INFINI Labs
极限科技，让搜索更简单

ES 与向量技术融合后的场景

语义搜索

通过将查询和文档映射到相同的向量空间，我们可以计算它们之间的相似性，以实现更精确的搜索结果。例如，当用户输入“苹果”时，系统不仅可以找到包含“苹果”这个词的文档，而且还可以找到与“苹果”语义相近的文档，如“iPhone”或“Macbook”。

推荐系统

在推荐系统中，我们可以利用用户的历史行为来学习用户的兴趣表示，然后用这个表示来检索与用户兴趣相似的项目。例如，如果一个用户经常搜索“Python 编程”相关的内容，那么我们就可以将这个用户的兴趣表示为“Python 编程”的向量，然后用这个向量来检索其他相关的内容，以实现个性化推荐。

内容聚类

向量表示可以用于测量文档之间的语义相似性，从而实现内容聚类。例如，新闻网站可以利用这个技术来找出关于同一主题的新闻文章，以便于用户阅读。

INFINI Labs
极限科技，让搜索更简单

如何利用向量提升 ES 搜索能力

使用通用句子编码 Universal Sentence Encoder（USE）

通用性	训练方法	实用性
USE可以在多种NLP任务上表现良好，如文本分类、语义、多语言支持。	采用了Transformer结构和深度平均网络（DAN）。Transformer结构可以捕捉句子内部复杂的上下文关系，而深度平均网络通过对句子中的所有单词向量进行平均来获取句子的表示，这使得USE模型在处理长句子时具有较高的效率。	预训练的模型可以直接用于各种任务，方便用户使用。

INFINI Labs
极限科技，让搜索更简单

如何利用向量提升 ES 搜索能力

加载 Universal Sentence Encoder 多语言版模型

```
model = hub.load("https://tfhub.dev/google/universal-sentence-encoder-multilingual/3")


embeddings = model(sentences)

return np.array(embeddings).tolist()
```

INFINI Labs
极限科技，让搜索更简单

如何利用向量提升 ES 搜索能力

INFINI Console 操作 ES 可无缝

对跨版本的集群进行管理

```
PUT knn-test
{
  "mappings": {
    "properties": {
      "title": {
        "type": "text"
      },
      "title_vec": {
        "type": "dense_vector",
        "dims": 512
      }
    }
  }
}
```

INFINI Labs
极限科技，让搜索更简单

如何利用向量提升 ES 搜索能力

导入新闻标题样本到 ES

```python
def main(sentences):
    embeddings = encode_sentences(sentences)
    for i, embedding in enumerate(embeddings):
        print(f"Sentence: {sentences[i]}")
        print(f"Embedding: {embedding}\n")
        document = {
            "title": sentences[i],
            "title_vec": embedding
        }
        es.index(index="knn-test", body=document)


if __name__ == "__main__":
    sentences = ["马斯克：地球上的机器人数量终会超过人类，应重视对AI的监管",
                 "7年内超级AI将问世！OpenAI宣布：20%算力投入，4年内控制超级智能",
                 "考生走路甩手误将准考证扔河里 消防员跳河打捞",
                 "2024年美国总统大选：共和党参选人数上升， 选情趋白热化"]
    main(sentences)
```

INFINI Labs
极限科技，让搜索更简单

如何利用向量提升 ES 搜索能力

对一条社会新闻做向量查询

预期属于社会新闻

> 北京朝阳大悦城现砍人事件
> 警方通报嫌疑人被控制

INFINI Labs
极限科技，让搜索更简单

如何利用向量提升 ES 搜索能力

普通match query

返回结果不符合预期。

```
GET /knn-test/_search?pretty
{
  "query" : {
    "match": {
      "title": "北京朝阳大悦城现砍人事件 警方通报嫌疑人被控制"
    }
  }
}
```


```
},
"hits": {
  "total": {
    "value": 3,
    "relation": "eq"
  },
  "max_score": 2.635185,
  "hits": [
    {
      "_index": "knn-test",
      "_type": "_doc",
      "_id": "CFvUKYkBZVYjwAb-aEHB",
      "_score": 2.635185,
      "_source": {
        "title": "2024年美国总统大选：共和党参选人数上升， 选情趋白热化",
        "title_vec": [...]
      }
    },
    {
      "_index": "knn-test",
      "_type": "_doc",
      "_id": "BlvUKYkBZVYjwAb-aEGY",
      "_score": 2.3285627,
      "_source": {
        "title": "7年内超级AI将问世！OpenAI宣布：20%算力投入，4年内控制超级智能",
        "title_vec": [...]
      }
    },
    {
      "_index": "knn-test",
      "_type": "_doc",
      "_id": "BVvUKYkBZVYjwAb-aEF6",
      "_score": 1.8625023,
      "_source": {
        "title": "马斯克：地球上的机器人数量终会超过人类，应重视对AI的监管",
        "title_vec": [...]
      }
    }
  ]
}
}
```

INFINI Labs
极限科技，让搜索更简单

如何利用向量提升 ES 搜索能力

Elasticsearch 8.0以前

script_score

```
GET /knn-test/_search?pretty
{
  "query": {
    "script_score": {
      "query": {
        "match_all": {}
      },
      "script": {
        "source": "cosineSimilarity(params.query_vector, 'title_vec') + 1.0",
        "params": {
          "query_vector": [...]
        }
      }
    }
  }
}
```

INFINI Labs
极限科技，让搜索更简单

如何利用向量提升 ES 搜索能力

返回的结果排序和预期一致

INFINI Labs
极限科技，让搜索更简单

如何利用向量提升 ES 搜索能力

8.0 版本以上 增加专门的 kNN search

支持 Approximate kNN

dims: Can’t exceed 1024 for indexed vectors ("index": true),
or 2048 for non-indexed vectors.

index: 如果为true，可启用kNN搜索，默认为false

similarity: 用于kNN搜索的相似度算法，如果 index 为true，则必须

```
PUT knn-test
{
  "mappings": {
    "properties": {
      "title": {
        "type": "text"
      },
      "title_vec": {
        "type": "dense_vector",
        "dims": 512,
        "index": true,
        "similarity": "l2_norm"
      }
    }
  }
}
```

INFINI Labs
极限科技，让搜索更简单

如何利用向量提升 ES 搜索能力

8.0 版本以上导入数据后查询

num_candidates: 每个分片考虑的最近邻候选者的数量。不能超过10,000 增加 num_candidates 倾向于提高最终 k 个结果的准确性，但是会耗费更长的时间

k: 要返回的与查询向量最相近（即最相似）的向量的文档数量

```
GET knn-test/_search
{
  "knn": {
    "field": "title_vec",
    "query_vector": [...],
    "k": 10,
    "num_candidates": 100
  },
  "fields": [ "title" ]
}
```

DataFun.

谢谢大家！

400 139 9200

INFINI Labs

极限科技，让搜索更简单