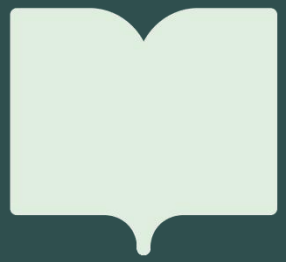
WSDM 2022 Best Paper

Learning Discrete Representations via Constrained Clustering for Effective and Efficient Dense Retrieval

CONTENTS



RepCONC

PART 01

当前困境

Research Motivation

Mainstream Trends

Core Challenges

Research Motivation-Dense Retrieval

BM25

核心机制：Term Matching
依赖精确的字面重合度进行检索。

痛点：Vocabulary Mismatch
无法理解语义。

Dense Retrieval

核心机制：Semantic Matching
将文本转为稠密向量，通过计算向量相似度匹配。

Query → Encoder → Embedding (向量)

Doc → Encoder → Embedding (向量)

最终：向量相似度计算

Research Motivation-Challenge

Memory Inefficiency

- ❖ 每个文档需存储高维稠密向量
- ❖ 总文档数量巨大，使得总存储需求巨大
- ❖ 计算量大，依赖昂贵GPU

Time Inefficiency

- ❖ brute-force计算，计算量大
- ❖ 高延迟导致实时性差，CPU难以满足需求
- ❖ 很难应用于实际情况

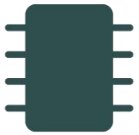
Challenge：如何在保证高召回率的前提下，降低硬件依赖并提高速度?

Core Challenges

Challenge：Effectiveness & Efficiency

Ranking Effectiveness

目标是保持甚至超越现有SOTA模型，确保召回结果的精准度与相关性。



Memory Efficiency

大幅降低索引文件体积，优化服务器内存占用，降低大规模部署的硬件成本。

Time Efficiency

在纯CPU环境下实现毫秒级检索响应，突破计算资源瓶颈，提升服务吞吐量。

PART 02

现有方案局限性

Mainstream Trends

方案一：无监督量化 (PQ, LSH)

核心思想：
用一个表示替代一群相似的表示。

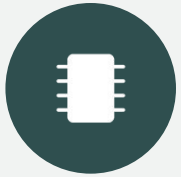
主要局限：
无监督学习导致量化过程与下游检索任务脱节，忽略了语义信息

方案二：有监督量化 (JPQ)

核心思想：
尝试将量化过程与模型训练相结合，引入监督信号以提升量化后的表征能力。

主要局限：
难以实现真正的端到端优化。例如JPQ算法中index assignment固定，仅优化query encoder和centroid embedding，极大限制了模型的最终性能上限。

无监督量化方案



阶段一：向量表示学习

- **方法：** 双编码器端到端训练，专注排序任务
- **目标：** 学习高质量的查询与文档向量表示
- **特点：** 此阶段完全不涉及量化操作

阶段二：无监督量化

- **方法：** 对已训练的文档向量单独做乘积量化(PQ)
- **目标：** 最小化向量与聚类中心的距离误差
- **特点：** 仅使用无监督信号，脱离检索任务

目标脱节

量化过程与检索任务优化目标完全解耦，缺乏针对性。

性能损失

量化误差会严重破坏原始向量的排序空间，导致召回率下降。

有监督量化方案：一种没有实施的思路

监督联合优化思路

核心思路：量化与训练深度耦合，允许索引分配动态更新，打破静态限制。

优化目标：让量化编码直接从检索任务的监督信号中学习，提升语义相关性。

实践中的两大核心挑战

“假语义区分”问题

模型利用非语义的虚假特征进行区分，导致泛化能力严重下降，检索结果不可靠。

梯度不稳定震荡

索引分配的离散性导致梯度剧烈震荡，训练过程极不稳定，难以收敛到全局最优解。

有监督量化方案：JPQ

Jointly Optimizing Query Encoder & PQ

核心思想：
沿用joint的思想，但是固定index assignment，从而解决梯度震荡问题，但是“假语义”问题仍没解决

Step 1: 固定索引
利用K-Means无监督预计算文档索引，训练期保持该分配不变。

Step 2: 联合训练
仅更新查询编码器与聚类中心参数，文档端冻结。

核心局限（RepCONC挑战1）：表达受限
比如一开始如果正样本和负样本在同一个cluster中就永远分不开了

RepCONC：想要语义精确，改变index，但是离散（Non-differentiability）

如果想要改变index：

解决量化操作不可微
使得可以反向传播优化参数

解决离散梯度震荡问题
使得结果能够收敛

PART 03

RepCONC

RepCONC 核心架构与训练思路

PQ + encoder联合训练

进行端到端联合训练，最大化表征精度，**不固定index**。

Constrained Clustering

语义表达尽可能准确+尽可能最大化区分度

IVF

降低候选集数量，进一步加快检索速度。

RepCONC

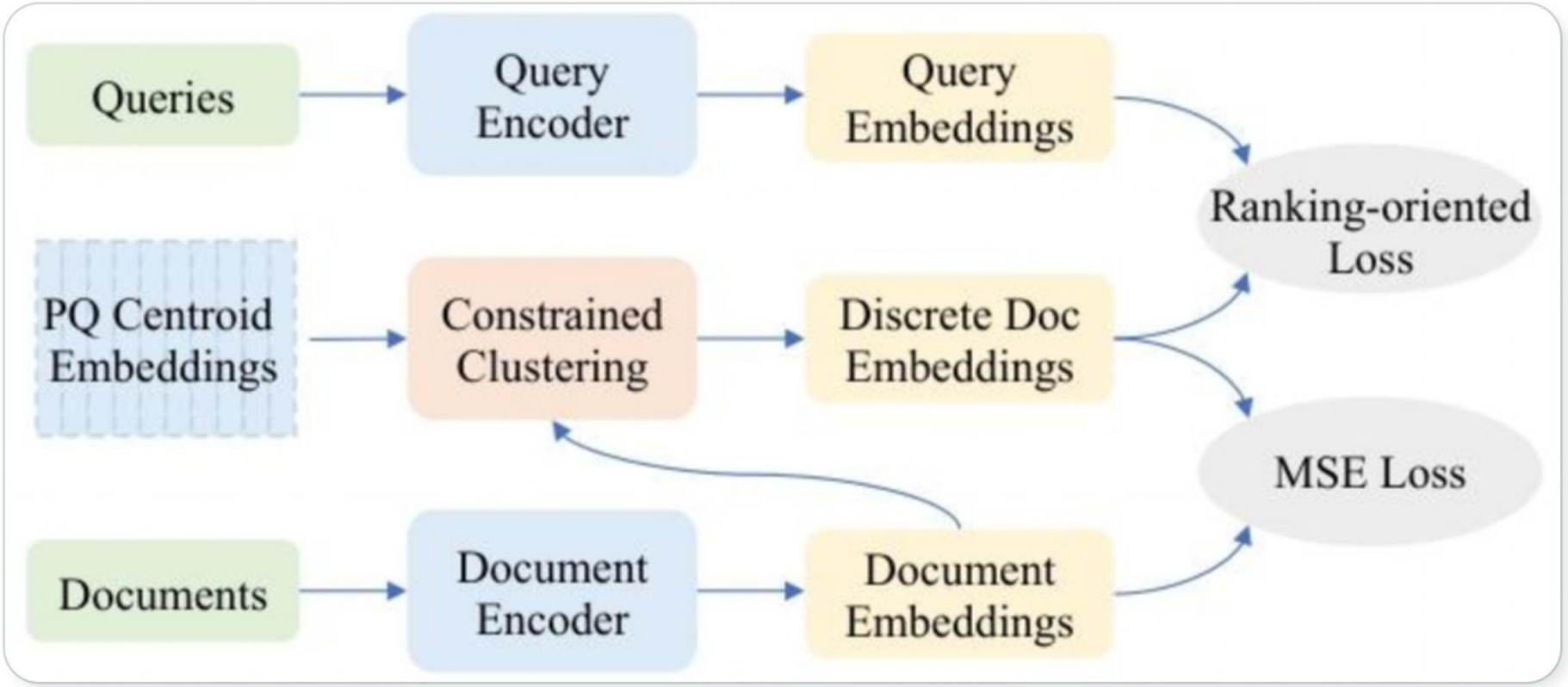


Figure 1: Training process of RepCONC.

RepCONC： Constrained Clustering 证

证明constrained clustering可以使得embedding间区分度最

$$\forall \boldsymbol{I} : P(\boldsymbol{\varphi}(d) = \boldsymbol{I}) = \frac{1}{|\mathcal{I}|} = \frac{1}{K^M} \tag{14}$$

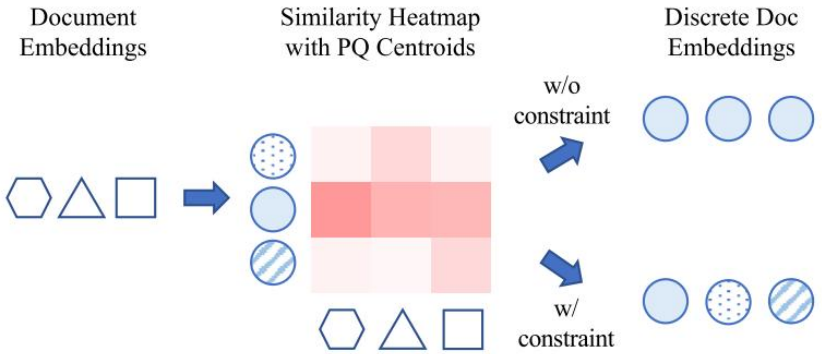


Figure 2: Illustration of Constrained Clustering. Darker colors in the heatmap indicate higher similarities (smaller distances). With the constraint, the discrete document embeddings are more diverse.

$$\forall i, j : P(\varphi_i(d) = j) = \sum_{\boldsymbol{I}:I_i=j} P(\boldsymbol{\varphi}(d) = \boldsymbol{I}) = \frac{|\{\boldsymbol{I} : I_i = j\}|}{K^M} = \frac{1}{K} \tag{15}$$

Eq14是Eq15的充分条件
在sub-vectors互相独立的情况下，Eq14和Eq15互为充分必要条件

RepCONC 语义

$$L = L_r + \lambda L_m \tag{8}$$

实现联合训练

$$L_r = -\log \frac{e^{\langle \boldsymbol{q}, \hat{\boldsymbol{d}}^+ \rangle}}{e^{\langle \boldsymbol{q}, \hat{\boldsymbol{d}}^+ \rangle} + \sum_{d^-} e^{\langle \boldsymbol{q}, \hat{\boldsymbol{d}}^- \rangle}} \tag{6}$$

任务目标实现

$$L_m = \|\boldsymbol{d} - \hat{\boldsymbol{d}}\|^2 \tag{7}$$

PQ训练，index不固定后，为了quantize后语义还精确

RepCONC　不能传播梯度问题

$$\frac{\partial L}{\partial \boldsymbol{d}} := \frac{\partial L_r}{\partial \hat{\boldsymbol{d}}} + \lambda \frac{\partial L_m}{\partial \boldsymbol{d}} \tag{9}$$

$$\begin{cases} \dfrac{\partial L}{\partial \hat{\boldsymbol{d}}} = \dfrac{\partial L_r}{\partial \hat{\boldsymbol{d}}} + \lambda \dfrac{\partial L_m}{\partial \hat{\boldsymbol{d}}} \\ \dfrac{\partial L}{\partial \boldsymbol{c}_{i,j}} = 1_{\varphi_i(d)=j} \cdot \dfrac{\partial L}{\partial \hat{\boldsymbol{d}}} \end{cases} \tag{10}$$

RepCONC 梯度震荡问题

Lm中的centroid每个子空间分别独立选择

$$\varphi_i(d) = \arg\max_j q(j|\boldsymbol{d}_i) \tag{17}$$

$$\forall i : \min_q \sum_{d\in\mathcal{B}} \sum_{j=1}^{K} q(j|\boldsymbol{d}_i)\|\boldsymbol{c}_{i,j} - \boldsymbol{d}_i\|^2$$

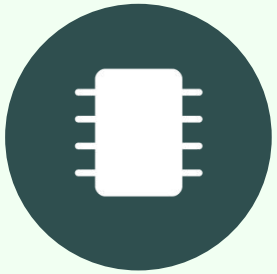
$$\text{subject to } \forall d : \sum_{j=1}^{K} q(j|\boldsymbol{d}_i) = 1 \text{ and } \forall j : \sum_{d\in\mathcal{B}} q(j|\boldsymbol{d}_i) = \frac{|\mathcal{B}|}{K} \tag{19}$$

使用Sinkhorn-Knopp algorithm

PART 04

优化方法

问题：带constrained cluster的 NP 问题



NP 难问题

离散+均匀约束

训练速度慢

训练不出来，因而不能实际应用

解决方案：转化为Optimal Transport

01. 离散变连续
将离散的 0/1 分配决策，松弛为连续的概率分布 q(j | d_i)

02. Cost Matrix
将量化误差 ||c_{i,j} - d_i||² 视为“运输成本”，量化分配决策的代价。

03. OT Problem
在满足均匀分布约束的前提下，最小化总体“运输成本”，实现高效求解。

解决方案

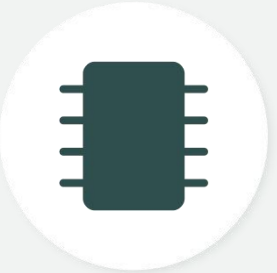
$$\varphi_i(d) = \arg\max_j q(j|\boldsymbol{d}_i) \tag{17}$$

$$\forall i : \min_q \sum_{d\in\mathcal{B}} \sum_{j=1}^{K} q(j|\boldsymbol{d}_i)\|\boldsymbol{c}_{i,j} - \boldsymbol{d}_i\|^2$$
$$\text{subject to } \forall d : \sum_{j=1}^{K} q(j|\boldsymbol{d}_i) = 1 \text{ and } \forall j : \sum_{d\in\mathcal{B}} q(j|\boldsymbol{d}_i) = \frac{|\mathcal{B}|}{K} \tag{19}$$

使用Sinkhorn-Knopp algorithm

Sinkhorn-Knopp 算法

- 将离散问题转化为连续问题
- 优化的时候梯度方向更稳定



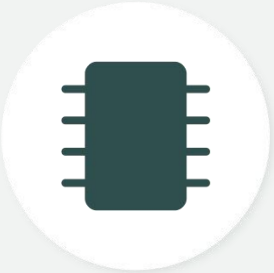
batch 替代全空间

- 所有Doc一起计算计算量太大
- 每个batch分别满足均匀条件，最终直接用最近距离得到index

其他

用Eq15代替14

- 需满足独立子空间条件
- 用边缘分布替代联合分布，减小计算量



Two-Stage Negative Sampling

- 第一阶段进行使用static hard negatives
- 第二阶段使用dynamic hard negatives

其他

IVF

- 对centroid embedding使用k-means
- 每次只检索少量cluster

PART 05

实验

Datasets & Metrics

Dataset

Dataset: MS MARCO
TREC 2019 DL Track

Task:

- Passage Ranking (8.8M passages)
- Document Ranking (3.2M documents)

Setting: Full-corpus

Metrics

- MRR@10 / MRR@100
- Recall@100 / NDCG@10

Baselines

有压缩

无监督方法 (Unsupervised)
- PQ / ScaNN / ITQ+LSH / OPQ

有监督方法 (Supervised)
- DPQ / JPQ

无压缩

传统与稠密检索
- 传统：BM25及变体
- 稠密：ANCE / ADORE

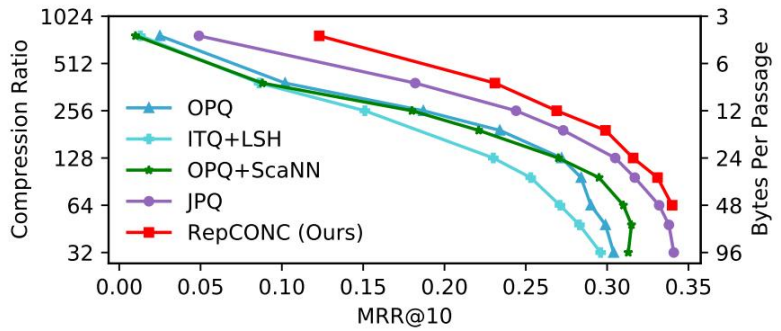
复杂端到端模型
- ColBERT / COIL

RQ1: 与压缩方法的比较

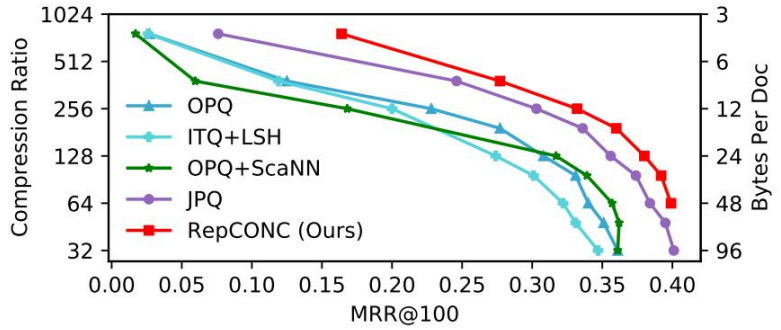
各压缩比下性能均最佳

高压缩比优势显著

RepCONC 兼顾了压缩与检索精度



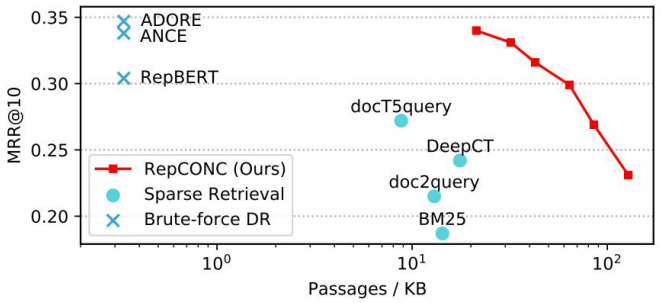
(a) MS MARCO Passage



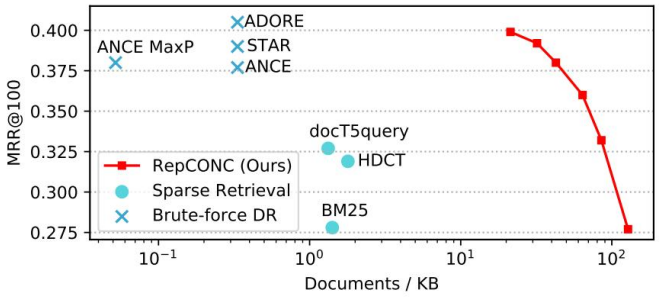
(b) MS MARCO Document

Figure 3: Comparison with compression methods. Up and right is better.

RQ2: 与未压缩模型的比较

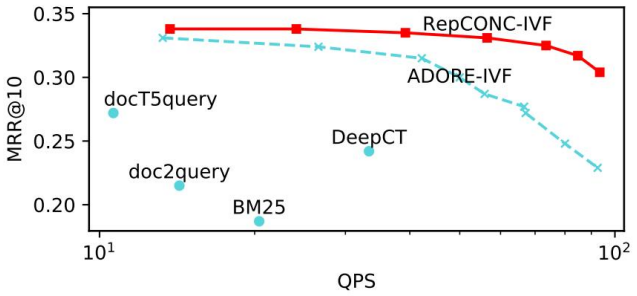


(a) MS MARCO Passage

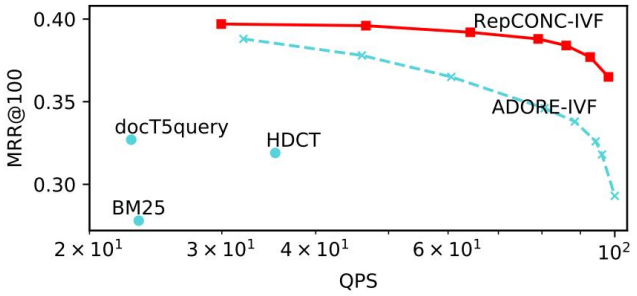


(b) MS MARCO Document

Figure 4: Comparison with first-stage retrieval models in terms of effectiveness-memory trade-off. Up and right is better. The x-axis indicates the average number of passages/documents stored in 1 kilobyte.



(a) MS MARCO Passage



(b) MS MARCO Document

Figure 5: Comparison with first-stage retrieval models in terms of effectiveness-latency trade-off. Up and right is better. The search is performed on CPU with one thread. QPS stands for 'query per second'.

空间效率高

在MRR接近最佳DR模型的前提下，极大缩减了索引体积。

卓越的检索性能

引入IVF索引后，RepCONC在CPU上的QPS远超同类模型。

RQ3: Ablation Study

Table 2: Ablation study on MSMARCO Passage Ranking dataset. BPP stands for 'bytes per passage'

Models	MRR@10 BPP:16	MRR@10 BPP:48
Baselines		
DPQ [4, 38]	0.244	0.305
JPQ [35]	0.273	0.332
RepCONC		
OPQ [11]	0.234	0.290
+ Clustering	0.275	0.332
+ Constraint	0.284	0.337
+ Dynamic Neg	0.294	0.340

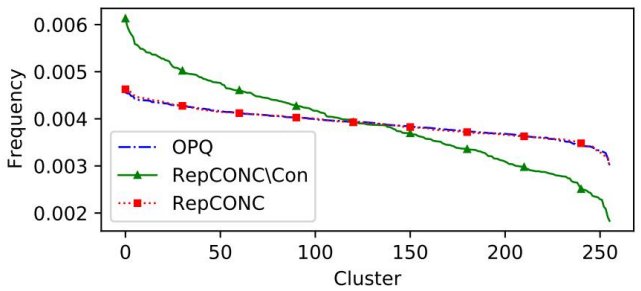


Figure 7: Cluster distribution on MS MARCO Passage Ranking. Clusters are sorted by the assigned frequency. Distributions across different sub-vector blocks are averaged. RepCONC\Con indicates RepCONC without constraint.

约束有效性验证

消融实验数据表明，加入均匀聚类约束（+Constraint）后，模型效果指标提升。

聚类分布更优

可视化结果清晰展示，带约束的RepCONC生成了更均衡的聚类分布，证明了约束机制的有效性。