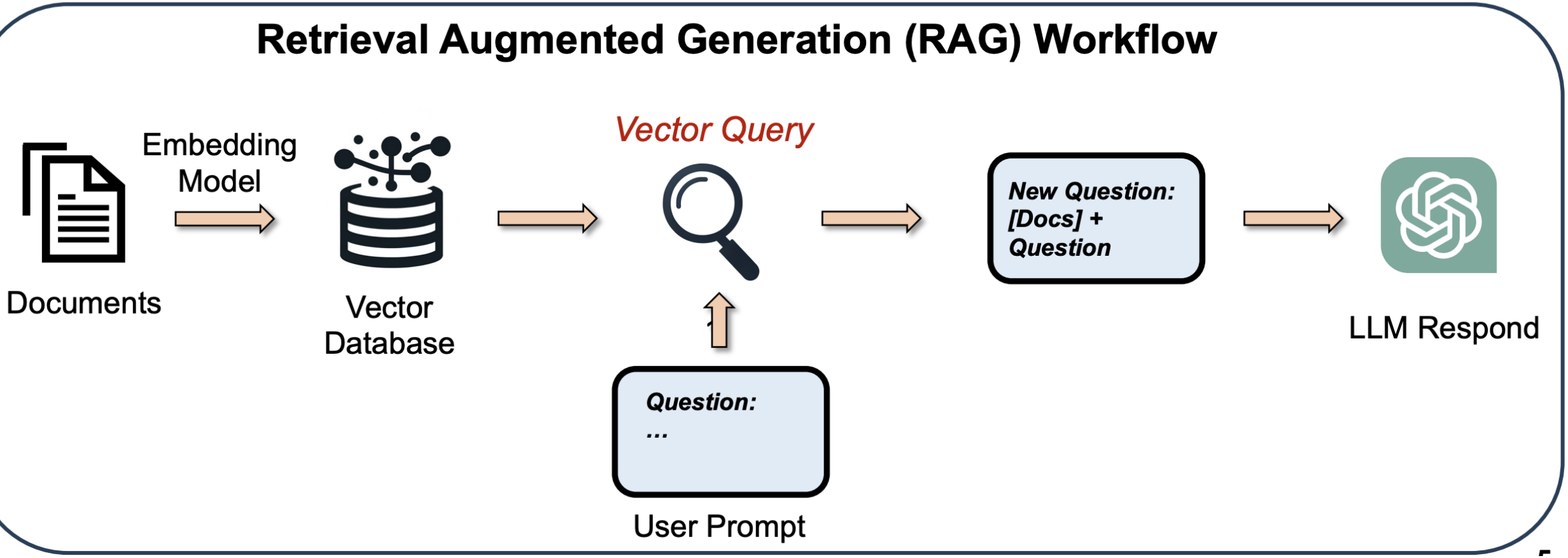
南京大学
NANJING UNIVERSITY

向量数据库

王智彬

背景介绍

Recap: RAG


Retrieval Augmented Generation (RAG) Workflow
Documents
Embedding Model
Vector Database
Vector Query
Question:
...
User Prompt
New Question:
[Docs] +
Question
LLM Respond

向量近似近邻搜索（ANNS）

- Meta的Faiss是一个用于高效相似性搜索和密集向量聚类的库。它包含搜索任意大小的向量集的算法。它还包含用于评估和参数调整的支持代码。

Pinecone

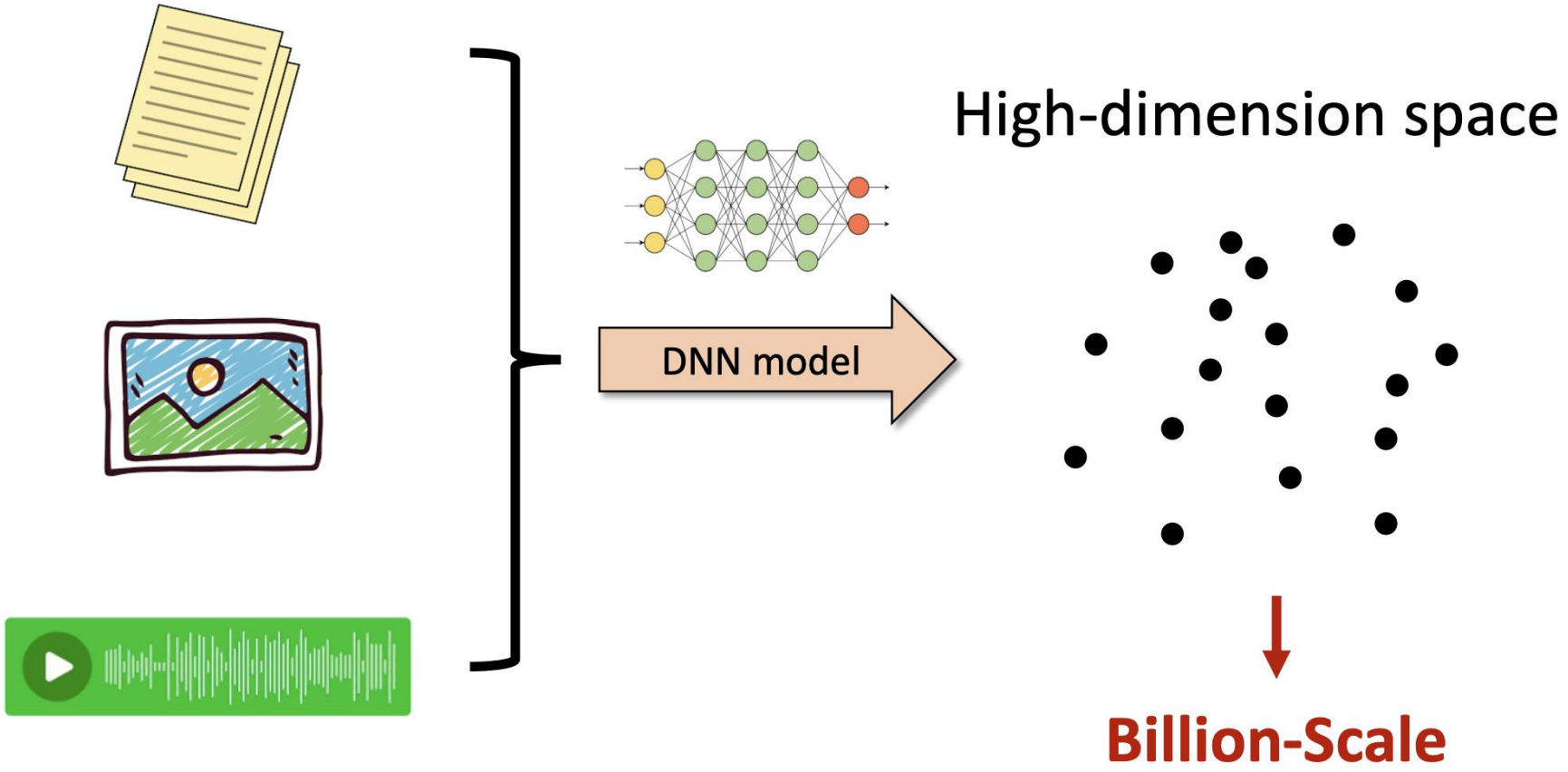
- Pinecone专为机器学习应用程序设计的矢量数据库。 它速度快、可扩展，并支持多种机器学习算法。建立在 Faiss 之上。

向量近似近邻搜索（ANNS）

- Milvus是一个开源矢量数据库，可以管理万亿矢量数据集，支持多种矢量搜索索引和内置过滤。

- Weaviate是一个开源向量数据库，允许存储数据对象和来自用户定义的 ML 模型的向量嵌入，并无缝扩展到数十亿个数据对象。

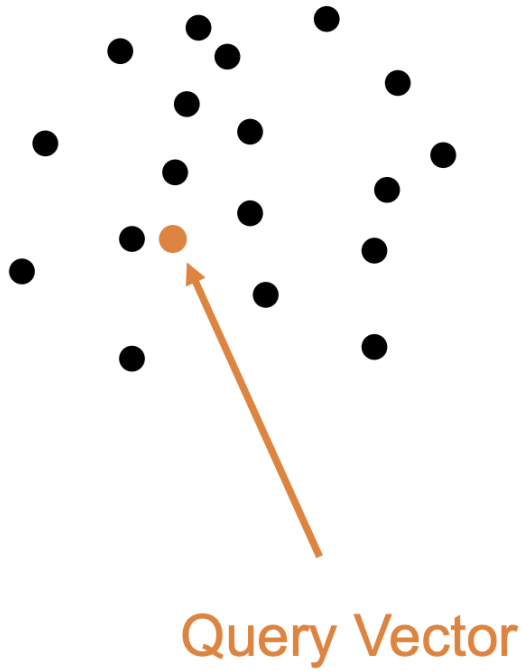
向量数据库



向量查询

- What is Vector Query ?

Given a query vector

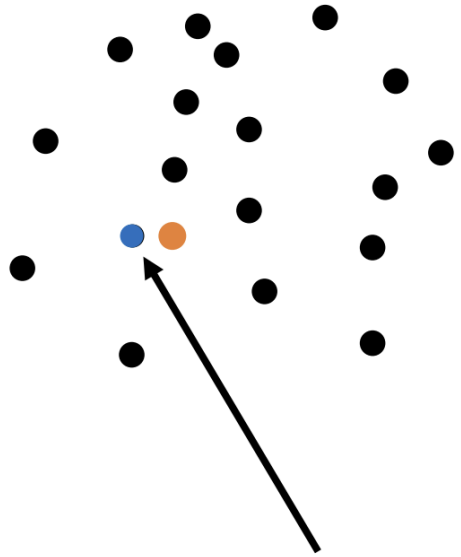


向量查询

- What is Vector Query ?

Given a query vector

Return Top-k Nearest vectors

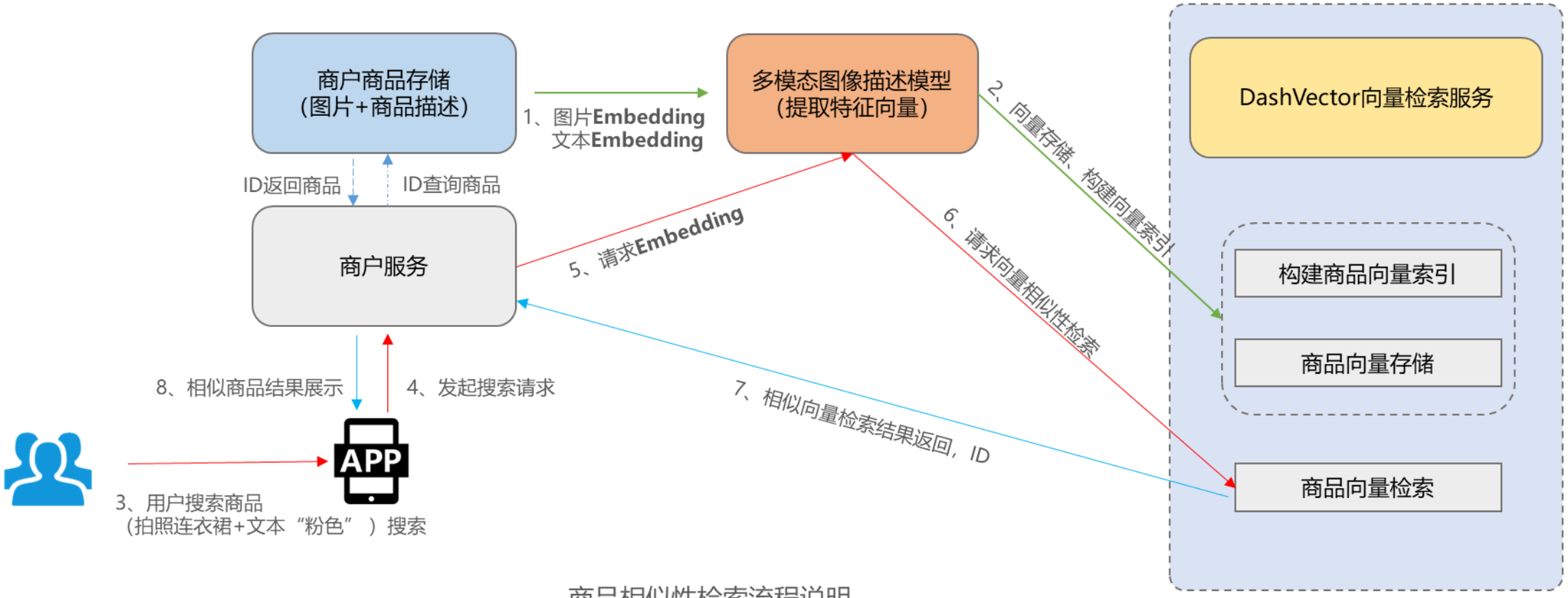


Top-1(k=1) Nearest Vector

更多向量数据库的应用

- 有同学能说说嘛

淘宝图片搜索类似商品



商品相似性检索流程说明

复杂度分析

- 给定n个向量，每个向量feature有d维，那knn的复杂度为
  - O(nd)
- 主流向量数据库大小
  - 1billion（十亿）到100billion（千亿）个向量
  - 每个向量128-1024维

近似计算：ANNS！

ANNS

目录

- ANNS的定义
- LSH算法
- IVF算法
- Graph-based算法

目录

ANNS

- Approximate Nearest Neighbor Search (ANNS)
- 向量数据库
  - 给定一个向量的点集S，包含n个点，每个点$\mathrm{x} \in R^d$。
  - 两个点x, y之间的距离函数定义为$\mathrm{D}(x, y)$
- ANNS查询
  - 给定一个查询$\mathrm{q} \in R^d$
  - ANNS要求返回一个点p
    - $P(D(p, q) \leq cr) > 1 - \delta$

ANNS

- Approximate Nearest Neighbor Search (ANNS)
- 向量数据库
  - 给定一个向量的点集S，包含n个点，每个点$\mathrm{x} \in R^d$。
  - 两个点x, y之间的距离函数定义为$\mathrm{Dis}(x, y)$
- ANNS查询
  - 给定一个查询$\mathrm{q} \in R^d$
  - ANNS要求返回一个点p
    - $Prob(Dis(p, q) \leq cr) > 1 - \delta$
    - 其中$\mathrm{r} = \min_x(D(p, x))$，就是实际中的最短距离
    - $1 - \delta$ 表示大概率是能够返回距离小于c的点

目录

- ANNS的定义
- LSH算法
- IVF算法
- Graph-based算法

LSH: Locality sensitive hashing

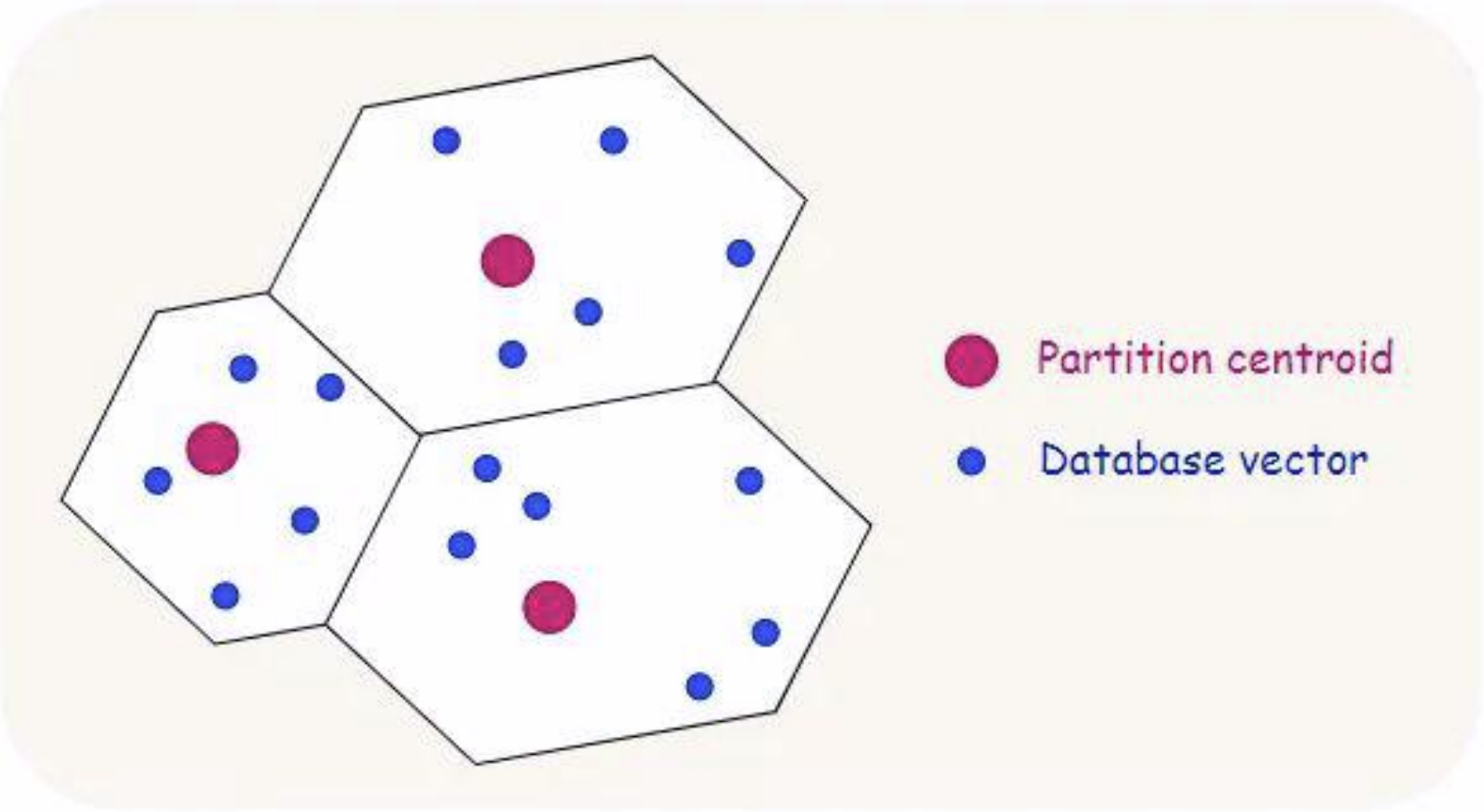
- 给定n个向量，每个向量feature有d维，那knn的复杂度为
  - O(nd)
  - **LSH关注如何降低d**
- 即给定一个向量$x \in R^d$，设计一个$\mathrm{LSH}(x) \in R^{d'}$
  - 对于欧氏距离$\mathrm{LSH}(x) = \left\lfloor \frac{xW+b}{w} \right\rfloor$，可以看成从高维到低维的投影
  - W是投影矩阵，维度为(d, d')，同时w可以降低数据的精度
- 一个简单的例子
  - 考虑一个二维的数据，我们把他映射到一维空间，一维空间的距离可以反应二维空间的距离，但会有精度损失
- 有理论保证！！！详见CS246W

目录

IVF：Inverted File

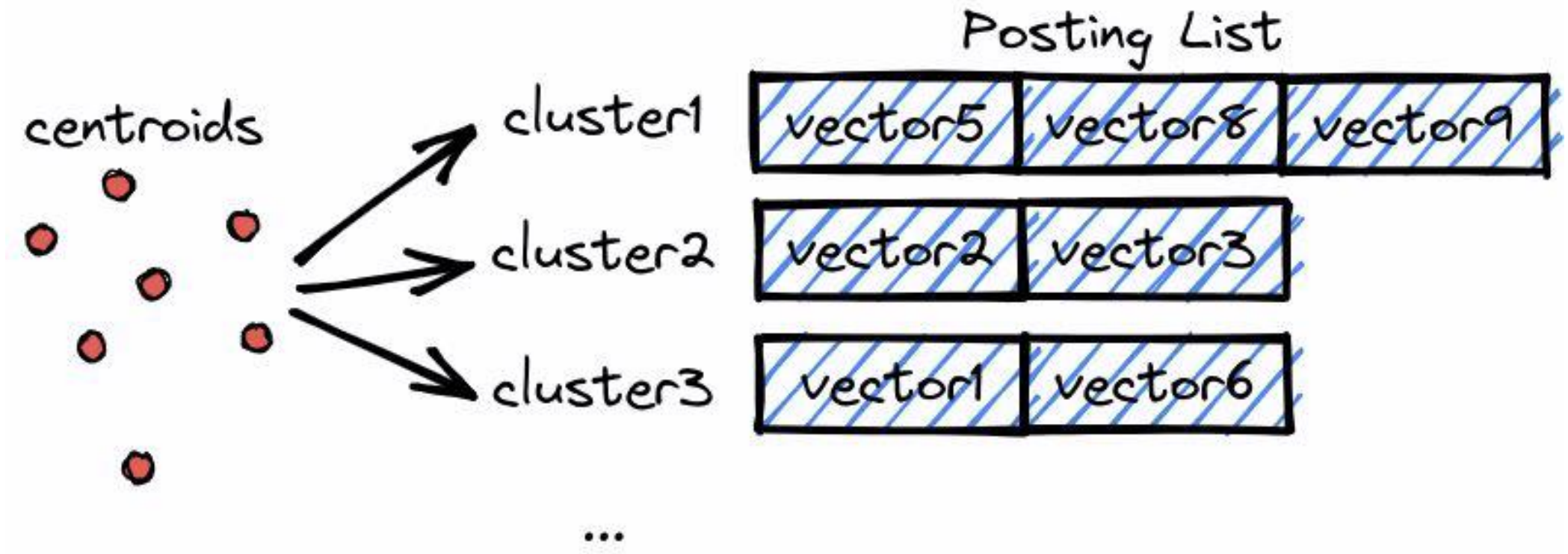
- 给定n个向量，每个向量feature有d维，那knn的复杂度为
  - O(nd)
  - **IVF关注如何降低n**
- **IVF过程**
  - 聚类阶段：使用算法（如 k-means）将所有向量划分为 k 个簇，每个簇由一个质心表示。
  - 索引构建：对于每个向量，找到其最近的质心，并将其分配到对应的簇中，形成倒排索引结构。
  - 查询过程：
    - 计算查询向量与所有质心之间的距离，选择距离最近的 m个质心
    - 仅在这 m 个簇中搜索，找到与查询向量最相似的向量。

IVF

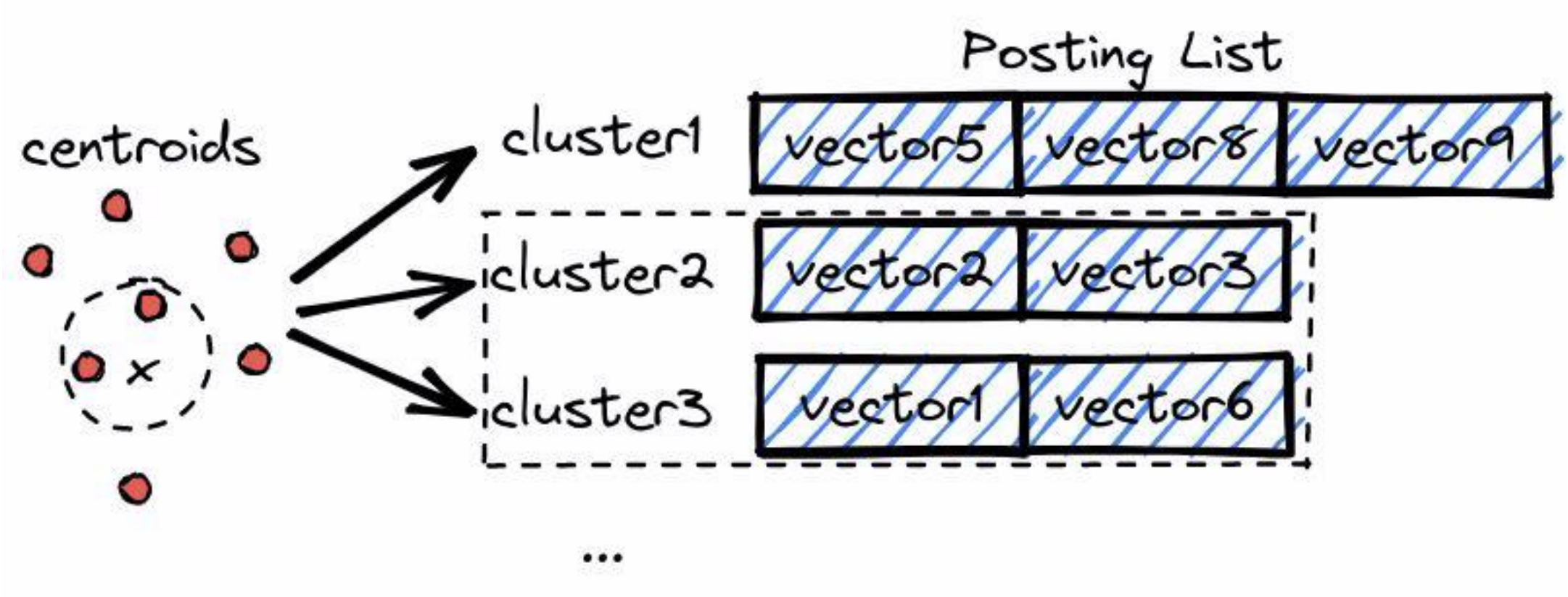


Credit: Pinecone

IVF

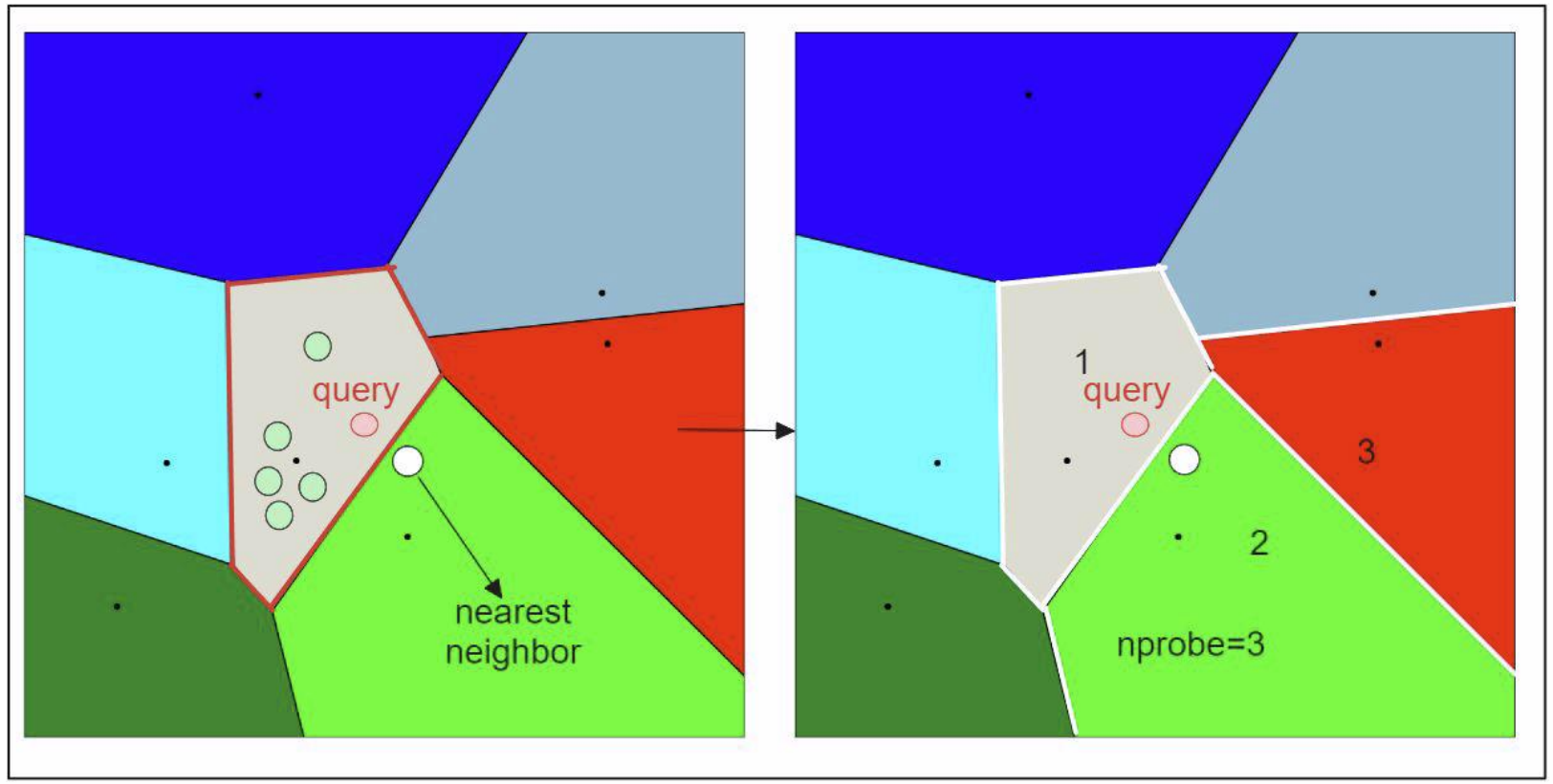

Posting List
centroids
cluster1
cluster2
cluster3
vector5
vector8
vector9
vector2
vector3
vector1
vector6
...

IVF


Posting List
centroids
cluster1
vector5
vector8
vector9
cluster2
vector2
vector3
cluster3
vector1
vector6
...

IVF：我们还有什么可以研究的

- 聚类算法的设计：速度快，聚类效果好
- 查找算法的设计



IVF：我们还有什么可以研究的

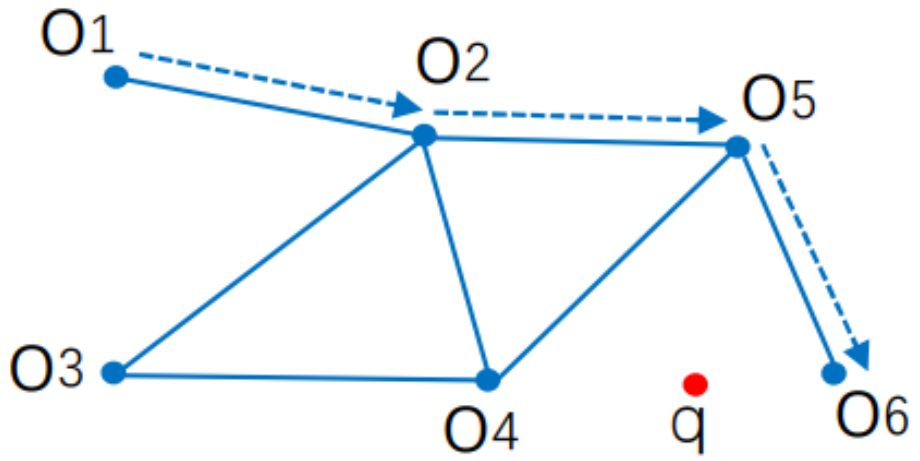
- 聚类算法的设计：速度快，聚类效果好
- 查找算法的设计
- 层次化内存下的向量数据库（系统设计）
  - 当内存放不下整个向量数据库，如何使用磁盘
  - NSDI 24的工作Fast Vector Query Processing for Large Datasets Beyond GPU Memory with Reordered Pipelining

目录

- ANNS的定义
- LSH算法
- IVF算法
- Graph-based算法

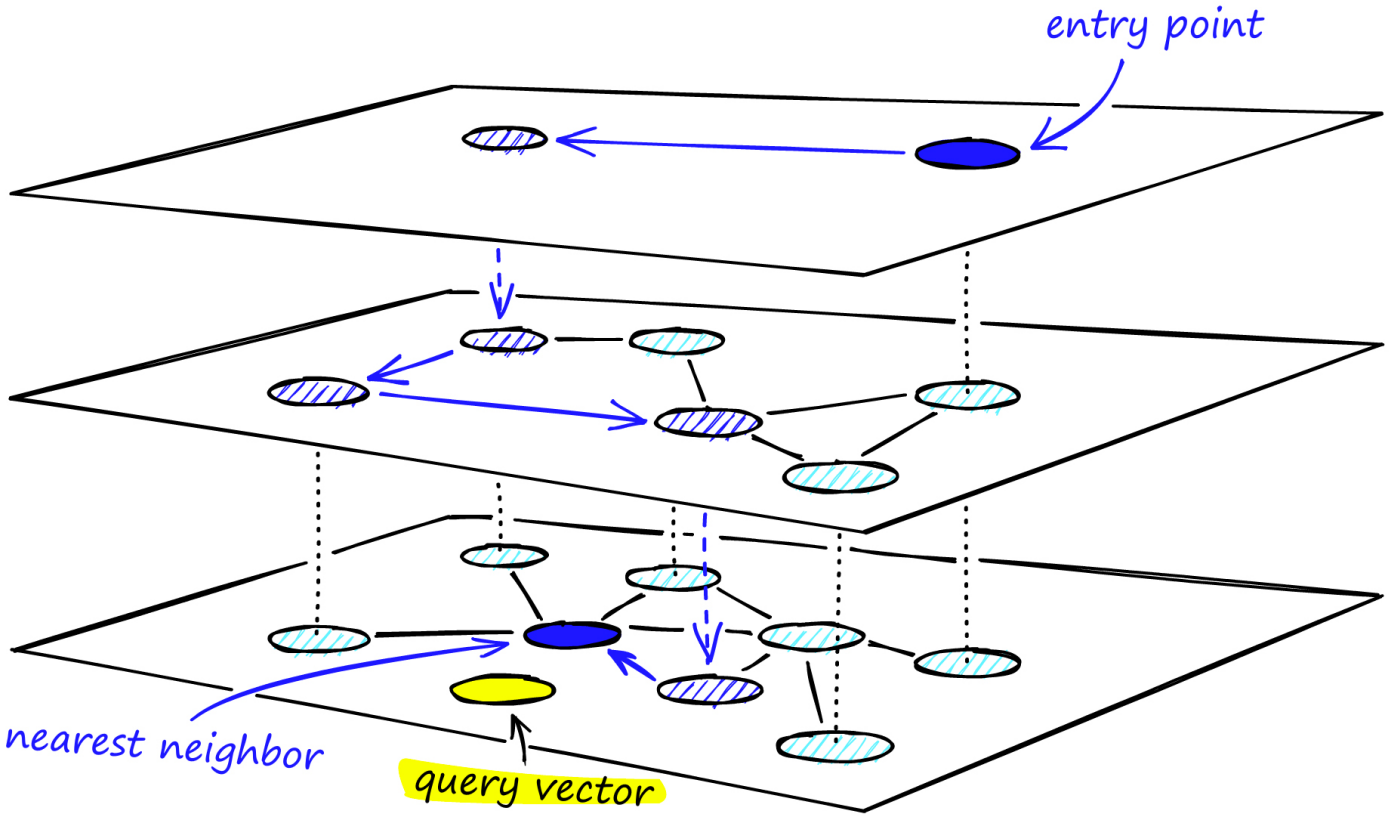
Graph-based ANNS

- 把向量的关系建模成一个graph
  - 每个向量是一个点
  - 距离近的向量连边
    - 例如使用KNN
- 搜索过程



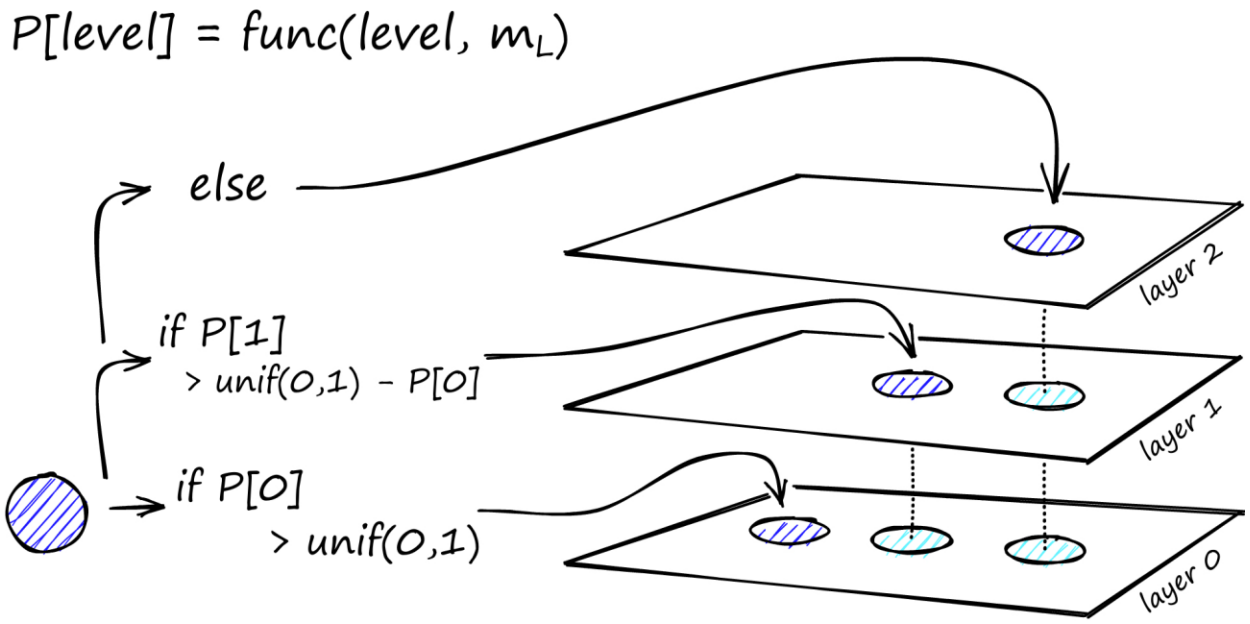
HNSW：Hierarchical Navigable Small Worl

- 分层的结构，类似IVF
  - 顶层较少
  - 底层较多
- 每一层都是用图表示
  - Small World Graph
  - 六度空间理论
    - 我们可以通过6个中间人联系到所有其他的人
  - 从任意一个节点出发，经过较少的跳跃就能到达目标节点



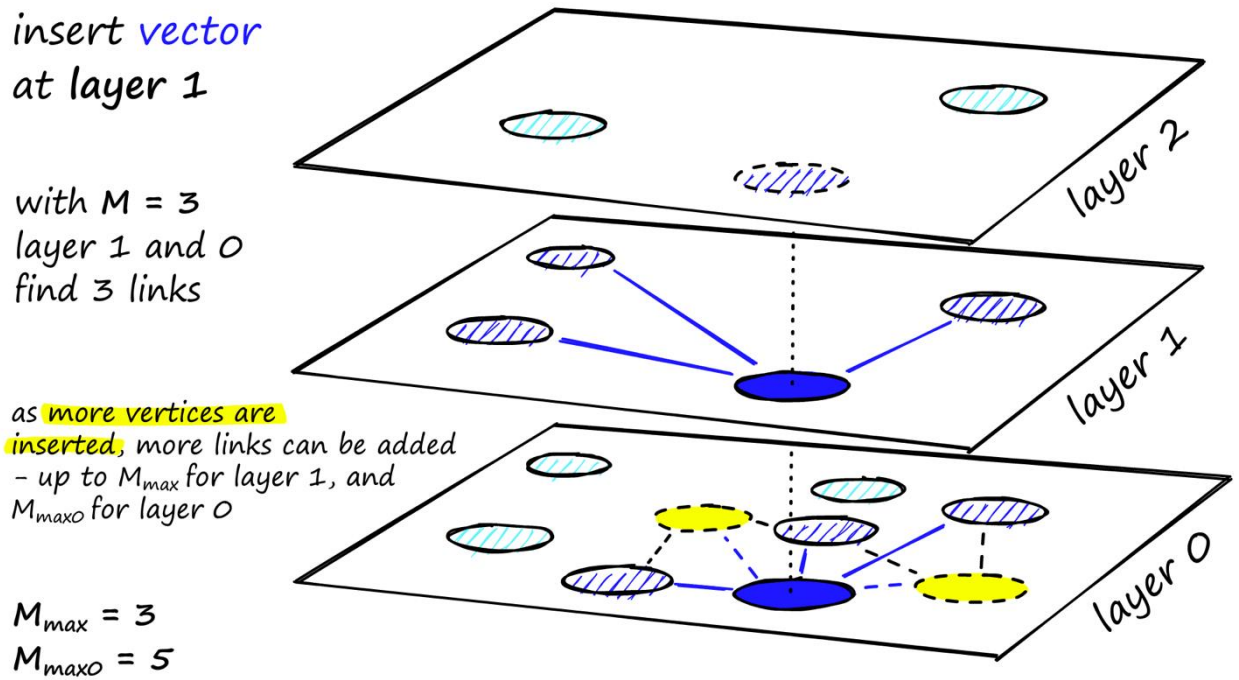
HNSW

- 在图构建过程中，向量逐个进行插入。层数由参数 L 表示。向量在给定层插入的概率由归一化的概率函数给出。



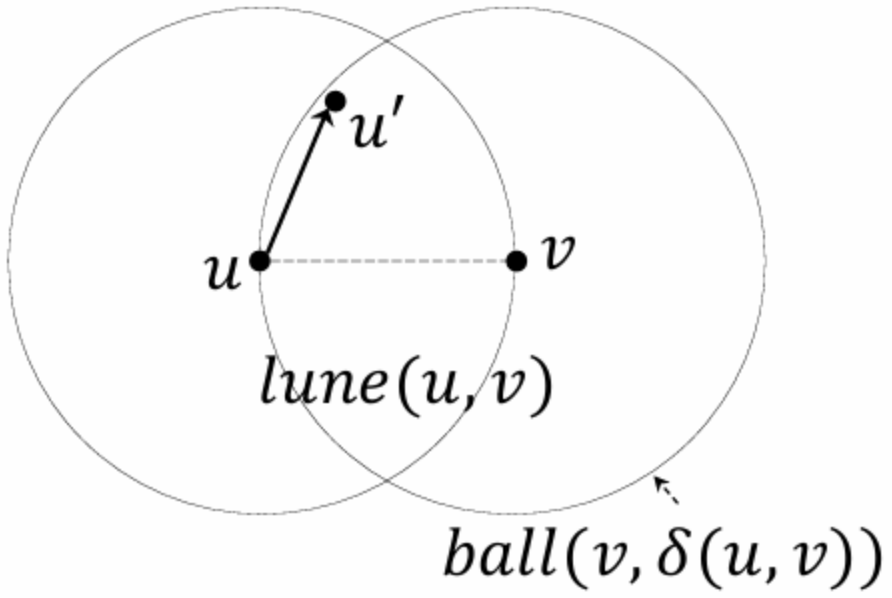
HNSW

- 在图构建过程中，向量逐个进行插入。层数由参数 L 表示。向量在给定层插入的概率由归一化的概率函数给出。
- 每次插入新节点，会建立M条边，M_{max}规定了邻居的最大个数。
- **NSW/HNSW没有提供严格的理论分析。**



RNG：Relative neighborhood graph

- 如果图中存在边(u,v)，那么图中就不会有点u'满足$dist(u,u') < dist(u,v)$和$dist(u',v) < dist(u,v)$。
- 可以减小图的规模



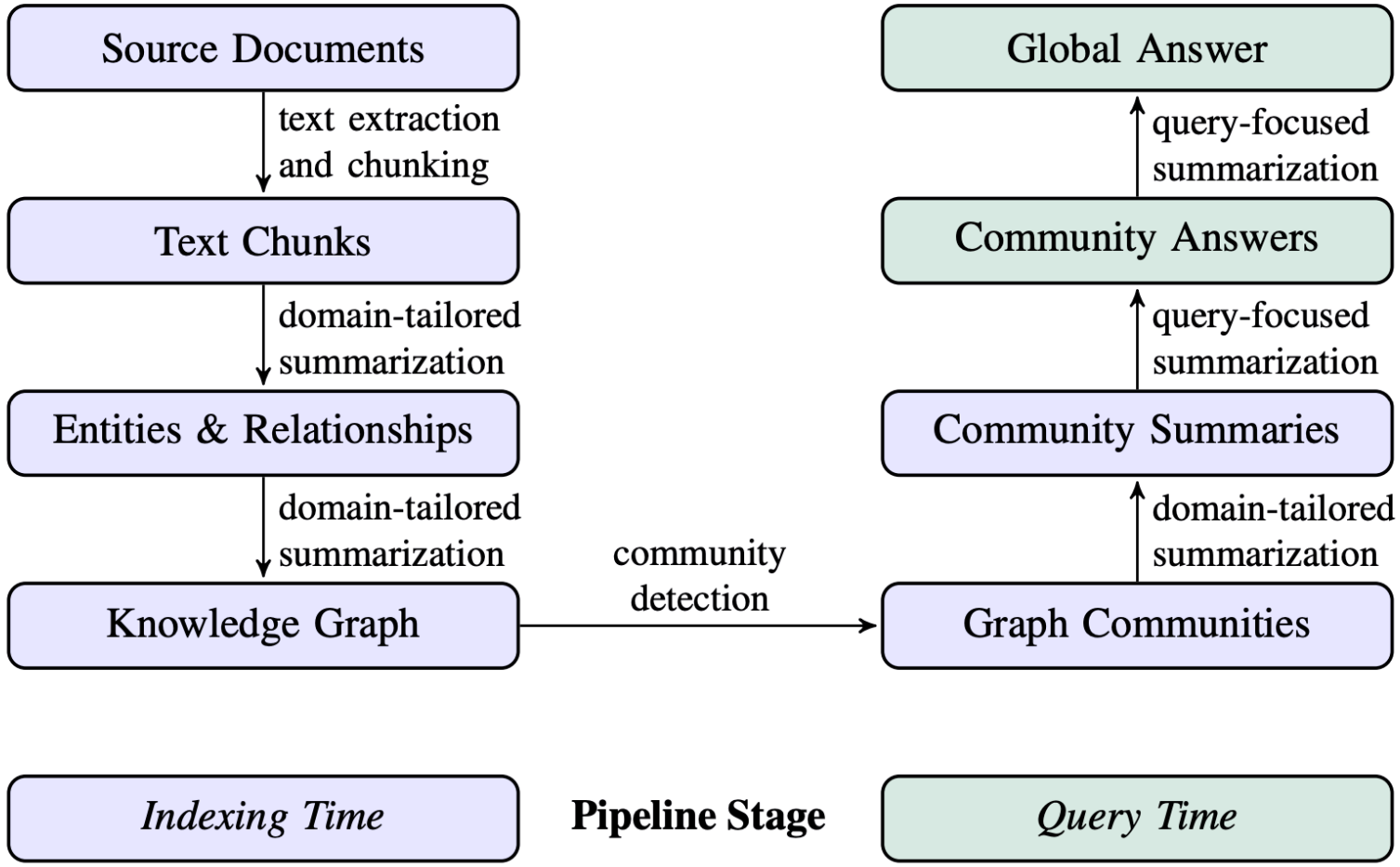
Graph-based ANNS还有什么可以研究的

- 基于内存的vector database，graph-based ANNS已经成为主流

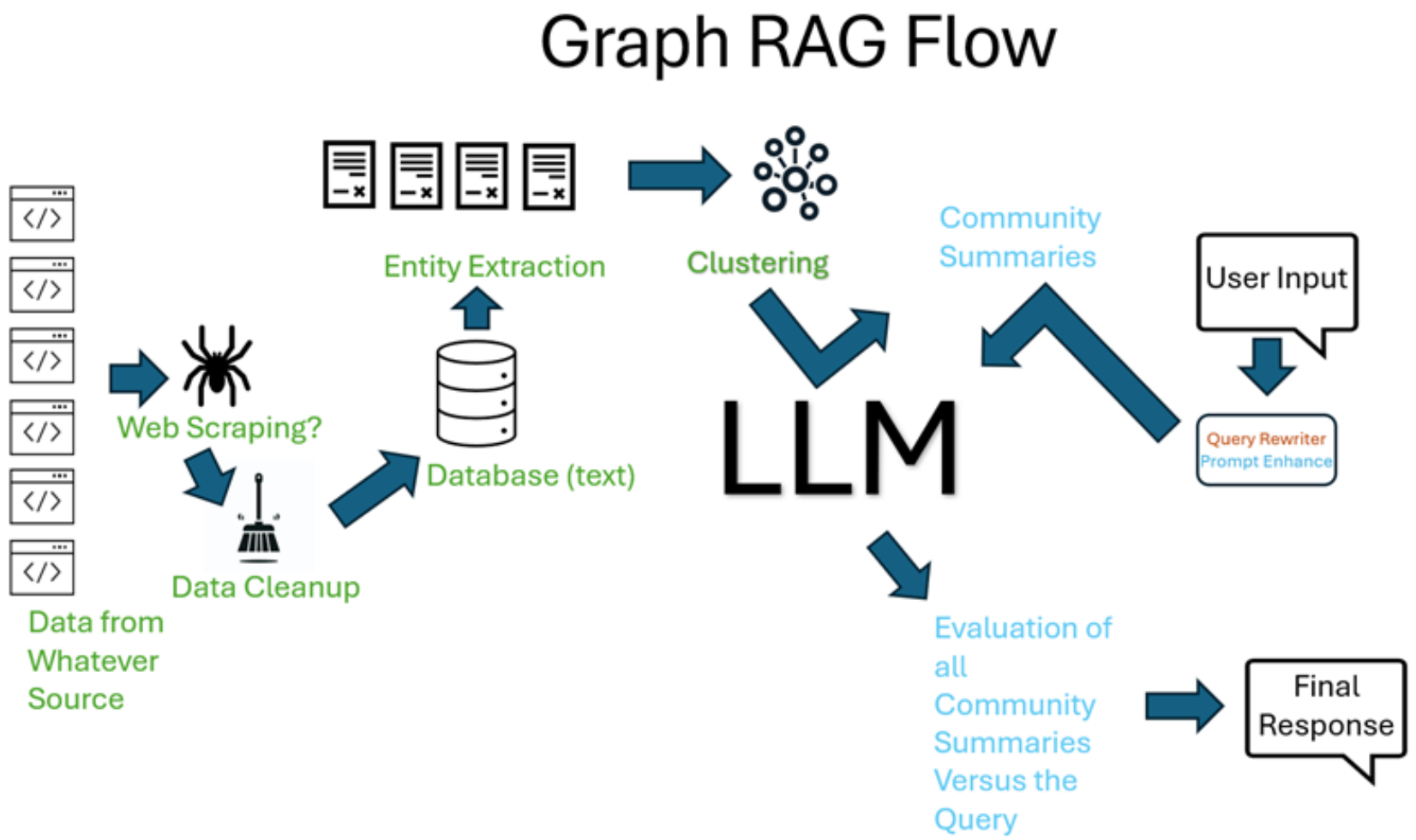
- 未来研究
  - 构建graph index的加速
  - 更好的graph index，减少查找量
  - 查找时的硬件加速
  - 查找过程中的剪枝

GraphRag

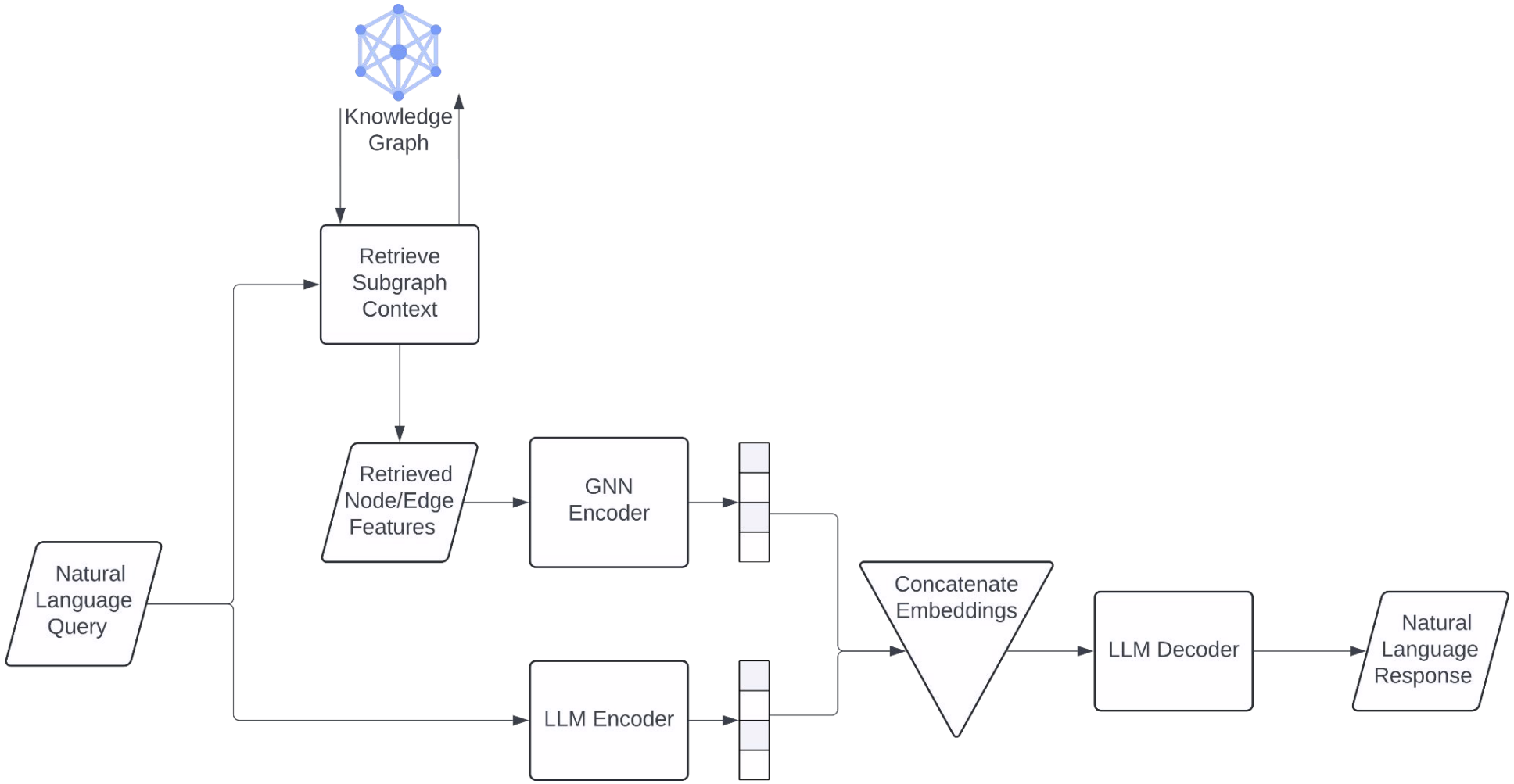
GraphRag


Source Documents
text extraction and chunking
Text Chunks
domain-tailored summarization
Entities & Relationships
domain-tailored summarization
Knowledge Graph
community detection
Graph Communities
domain-tailored summarization
Community Summaries
query-focused summarization
Community Answers
query-focused summarization
Global Answer
Indexing Time
Pipeline Stage
Query Time

GraphRag


Graph RAG Flow
Data from Whatever Source
Web Scraping?
Data Cleanup
Database (text)
Entity Extraction
Clustering
Community Summaries
LLM
User Input
Query Rewriter
Prompt Enhance
Evaluation of all Community Summaries Versus the Query
Final Response

GNN+RAG


Knowledge Graph
Retrieve Subgraph Context
Retrieved Node/Edge Features
GNN Encoder
Natural Language Query
LLM Encoder
Concatenate Embeddings
LLM Decoder
Natural Language Response

作业三

向量数据库加速

- 随着知识库的增大，向量数据库的查询成为了瓶颈
- 我们设计了数据集和题目
  - https://github.com/LTTMG/anns-problem
- 研究内容
  - 图索引的构建
  - 图索引的查找
  - IVF的构建
  - IVF的查找
- 研究方向
  - 算法：设计新算法，最小化计算量
  - 工程：在已有算法上，考虑并行，SIMD或者GPU，提升性能
  - 系统：在已有内存ANNS库上，考虑分布式或者外存场景，overlap传输和计算

GraphRag

- 目前大家都用vector database当做Ragas检索后端
- 能否改成使用使用knowledge graph作为知识库
  - 使用GraphRag
    - https://github.com/microsoft/graphrag
  - 或者GNN Rag
    - https://web.stanford.edu/class/cs224w/slides/18-LLMs+GNN.pdf

南京大学
NANJING UNIVERSITY

Thanks

Q&A

wzbwangzhibin@gmail.com