PostgreSQL中文社区

Pgvector构建私域知识库之存储和查询向量嵌入

Postgresql plug-in pgvector builds storage and query vector embedding of private domain knowledge base

分享人：周伦光

PostgreSQL中文社区

contents 目 录

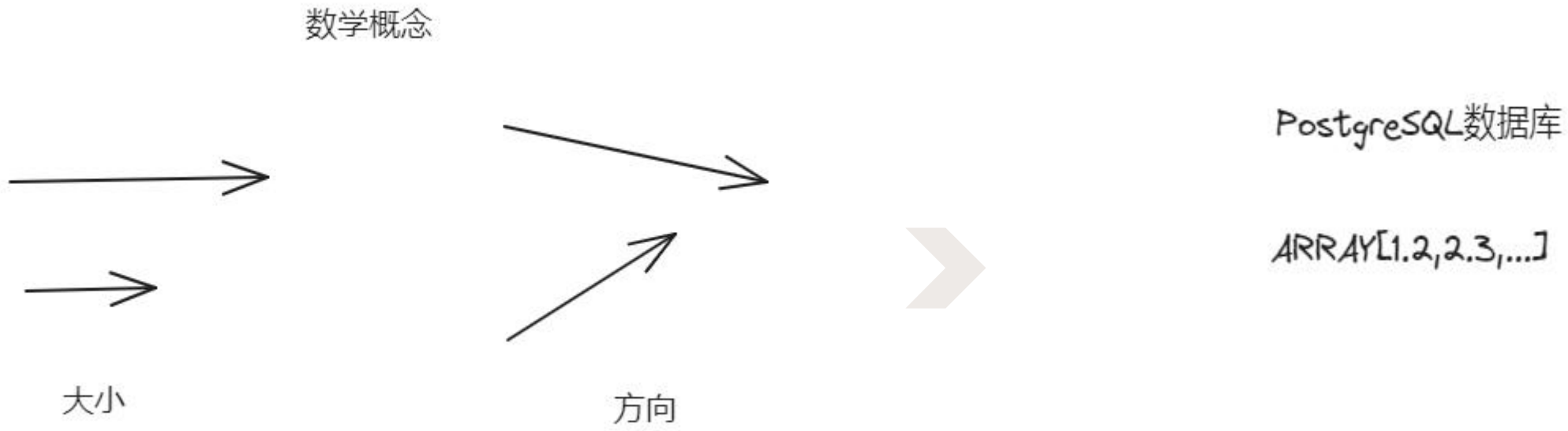
PostgreSQL中文社区

PART 01

原理分析

向量

PostgreSQL中文社区



相似性

给定两个属性向量， A 和B ,A_i和 B_i分别代表向量A和B的各分量，其余弦相似性θ由点积和向量长度给出

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

相似性范围从-1到1。-1意味着两个向量指向的方向正好截然相反，1表示它们的指向是完全相同的，0通常表示它们之间是独立的，而在这之间的值则表示中间的相似性或相异性

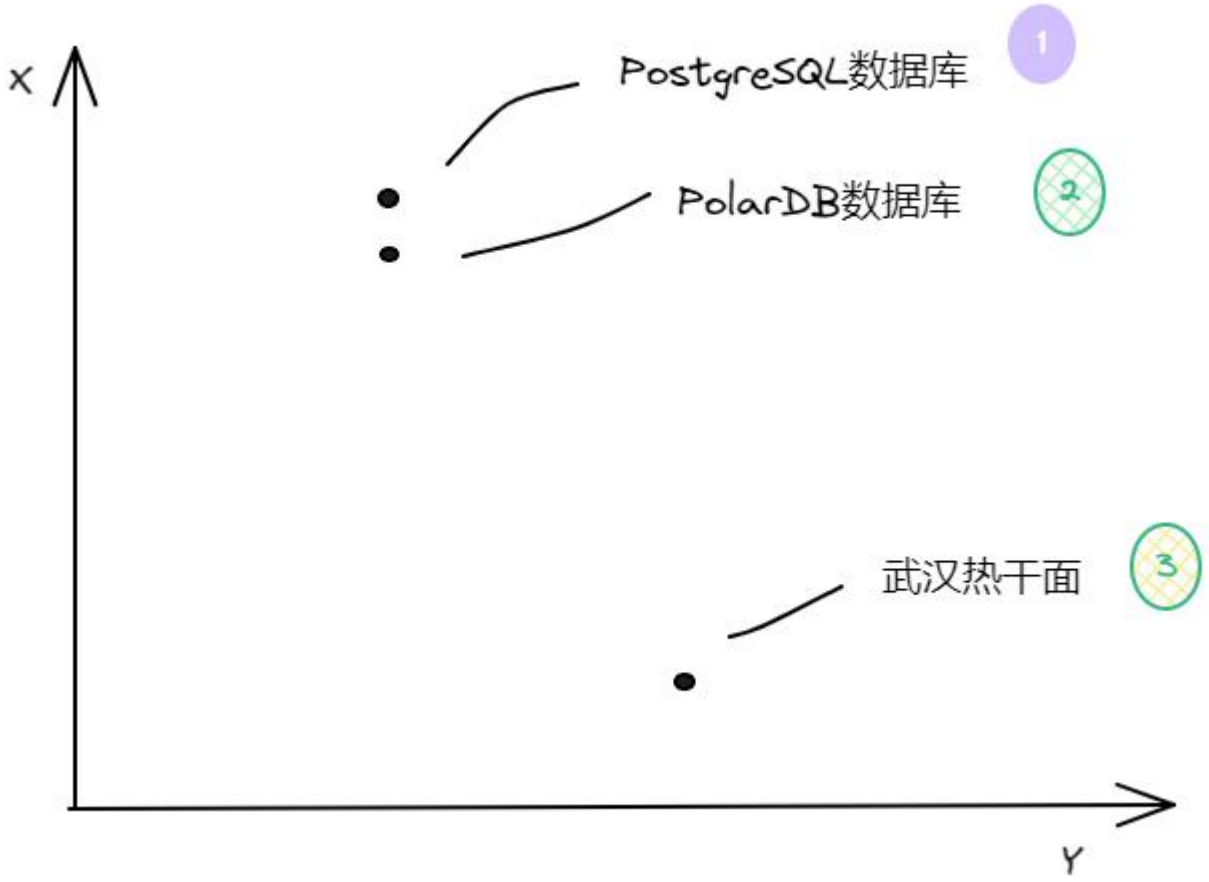
对于文本匹配，属性向量A 和B 通常是文档中的词频向量。余弦相似性，可以被看作是在比较过程中把文件长度正规化的方法

相似性

如果我们将短语画在2维图表上，维度：X 轴和 Y 轴

短语 1 和 2 将彼此靠近绘制，因为它们的语义相似

短语 3远离1和2，因为它们的语义相似过低



向量检索技术

PostgreSQL中文社区

NSW

关键是在构图过程中通过贪婪搜索算法记录下搜索最优路径，是一个类似于“交通枢纽”的无向图，这会导致某些顶点的出度激增

HNSW

对NSW的升级，基于图的方案，使用跳表结构代替NSW的链表结构通过空间换时间的方法将向量检索的复杂度从多重对数复杂度降至对数复杂度。

IVF_PQ

IVF可以将向量分解为若干个子空间，并且在每个子空间上使用倒排索引来加速搜索。PQ是一种有效的向量量化方法，可以将高维向量压缩为低维码本地，从而减少存储和计算成本

向量数据库原理

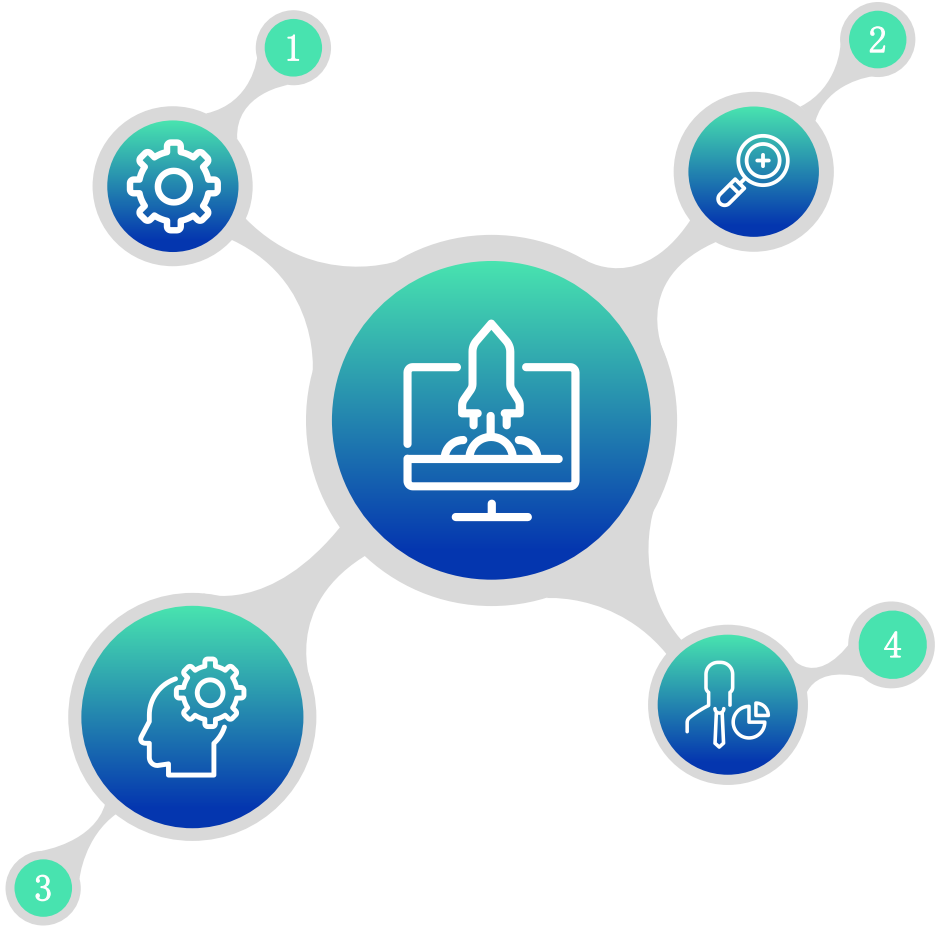
PostgreSQL中文社区

数据存储

将数据存储为向量形式，每个向量代表一个数据对象。向量的维度数取决于数据对象的特征数

向量索引

是一种数据结构，可以将向量数据按照一定的规则进行划分和组织，以便快速地进行查询和检索



相似度计算

查询操作主要是基于相似度计算计算该向量与数据库中所有向量的相似度，并返回相似度最高的几个向量作为查询结果

查询优化

为了提高查询效率，采用了一些查询优化技术，例如基于向量索引的查询优化、基于近似相似度计算的查询优化

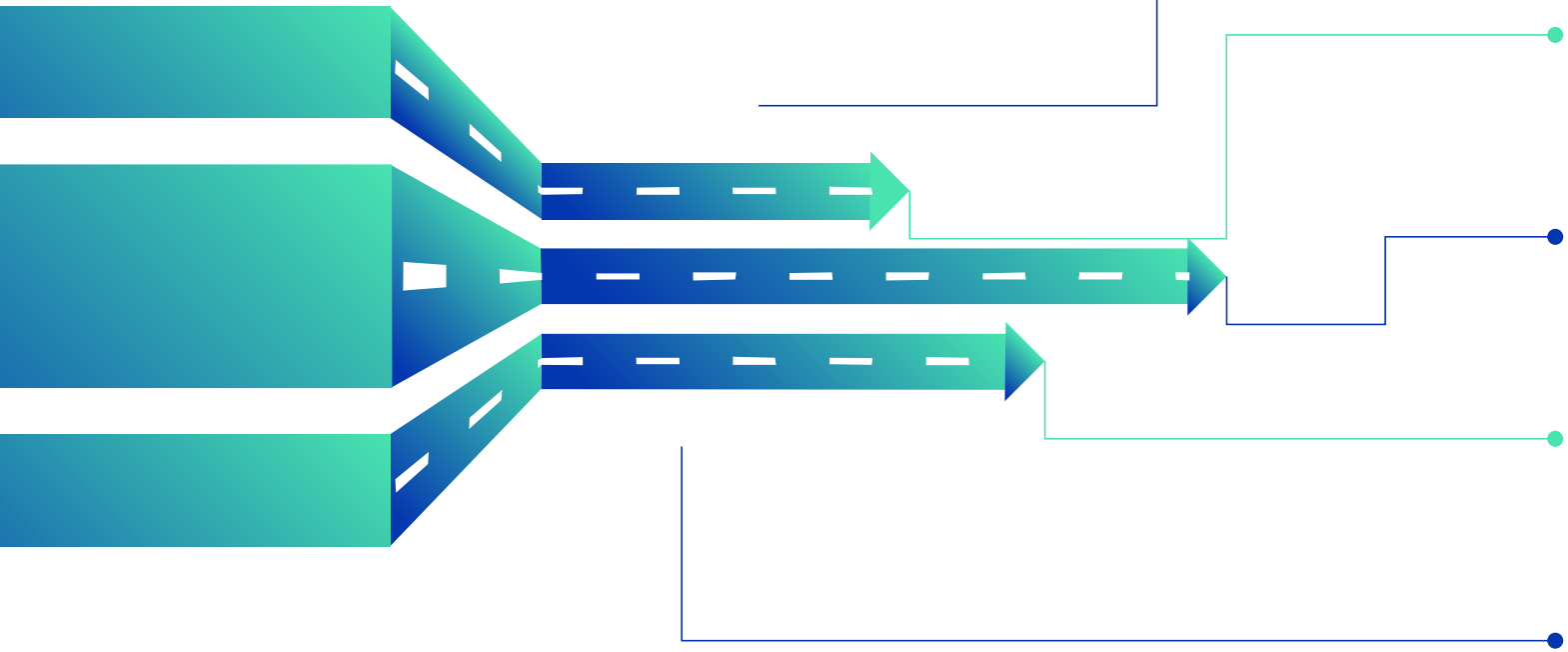
PostgreSQL中文社区

PART 02

应用实践

优势

PostgreSQL中文社区



高度集成

可以在现有的postgsql数据库中以插件的形式使用

支持多种距离度量

内置了对多种距离度量的支持，包括L2距离、余弦距离和内积。这种多功能性允许高度可定制的基于相似性的搜索和分析，以满足特定的应用需求

索引支持

提供了高效的索引选项，例如 k-最近邻 (k-NN) 搜索。因此，用户可以在保持高搜索准确性的同时实现快速查询执行

SQL

利用熟悉的 SQL 查询语法进行向量操作。这种方法简化了具有 SQL 知识和经验的用户采用矢量数据库的过程，并避免他们必须学习新的语言或系统

稳定性

pgvector 继承了postgresql的稳健性和安全性功能，与pg保持良好的版本和功能兼容

使用参考

PostgreSQL中文社区

01

向量类型

每个向量占用4 * dimensions + 8字节存储空间。每个元素都是一个单精度浮点数（类似于Postgres 中real不精确浮点类型），并且所有元素都必须是有限的（没有NaN, Infinity或-Infinity）。矢量最多可以有16,000 个维度。

02

向量运算符

操作符	描述
+	逐元素加法
-	逐元素减法
<-->	欧氏距离
<#>	负内积
<=>	余玄距离

03

向量函数

功能	描述
cosine_distance(vector, vector) → 双精度	余弦距离
inner_product(vector, vector) → 双精度	内积
l2_distance(vector, vector) → 双精度	欧氏距离
vector_dims(向量) → 整数	维数
vector_norm(vector) → 双精度	欧氏范数

04

聚合参数

功能	描述
avg(vector) → vector	平均值

存储向量

创建向量表:

```
CREATE TABLE items (id bigserial PRIMARY KEY, embedding vector(3));
```

插入向量:

```
INSERT INTO items (embedding) VALUES ('[1,2,3]'), ('[4,5,6]');

INSERT INTO items (id, embedding) VALUES (1, '[1,2,3]'), (2, '[4,5,6]')
    ON CONFLICT (id) DO UPDATE SET embedding = EXCLUDED.embedding;
```

更新向量:

```
UPDATE items SET embedding = '[1,2,3]' WHERE id = 1;

DELETE FROM items WHERE id = 1;
```

查询向量

获取向量的最近邻居:

SELECT * FROM items ORDER BY embedding <-> '[3,1,2]' LIMIT 5;

让最近的邻居排成一排:

SELECT * FROM items WHERE id != 1 ORDER BY embedding <-> (SELECT embedding FROM items WHERE id = 1) LIMIT 5;

获取距离:

SELECT embedding <-> '[3,1,2]' AS distance FROM items;

```
(base) [root@test.a8-33-3.dev.unp doc]# python search_vector.py  postgresql中文社区主席
--------------------------------------------------------------------------------
Score:  16.993154234394467
0000-b5e8-1887-9d4d-260.txt自1993年起，热心网友何伟平（网名laser，现国内数据库专家，”去哪儿“网站首席架构 师）开始独自翻
译并研究PostgreSQL十余年，网上现仍可随处搜索到何伟平当时所翻译的 PostgreSQL4.0---8.2等各版本的手册，尤其是那篇《Postgre
SQL的昨天、今天和明 天》在 网上广为流传，为广大中国IT开发人员了解PostgreSQL起到了重要的作用，在此我们也对何 伟平所作的
贡献表示由衷的敬意。 今天，随着众多使用和爱好PostgreSQL人员的加入，PostgreSQL在中国的应用呈快速发展之 势。
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
Score:  17.732505042883496
为进一步规范中文社区的发展，在今年2013杭州PostgreSQL用户大会上，经广大网友的 一致讨论，我们正式成立注册的组织---Postgre
SQL中国用户协会，并根据各位网友的特长 和爱好，成立不同的事务处理小组和核心的管理委员会，来共同努力更好地为广大网友服 务
，在中国大力推广这一优秀的数据库软件。 postgresql中文社区进入机制:原则上每年开展2次核心组新成员的引入投票工作；新成员加
 入前，需由其本人亲自陈述加入核心组的理由，由当前核心组成员在其陈述后，进行投票， 超过51%的成员同意，则同意新成员的加入
；
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
Score:  18.79688612871874
经同意加入后的新成员，即刻与原成员具有同 等投票权 postgresql主席选举:每届协会主席任期满后，由核心组成员会议投票选举新一
届的主席和 副主席；主席和副主席在任期内的约束等同于下述一般核心组成员的自动退出条件； postgresql中文社区主席:张文升 pos
tgresql中文社区常委:彭煜玮，周正中，朱贤文，唐成，姚延栋，姜明俊，张文升 第 1 页
--------------------------------------------------------------------------------
```

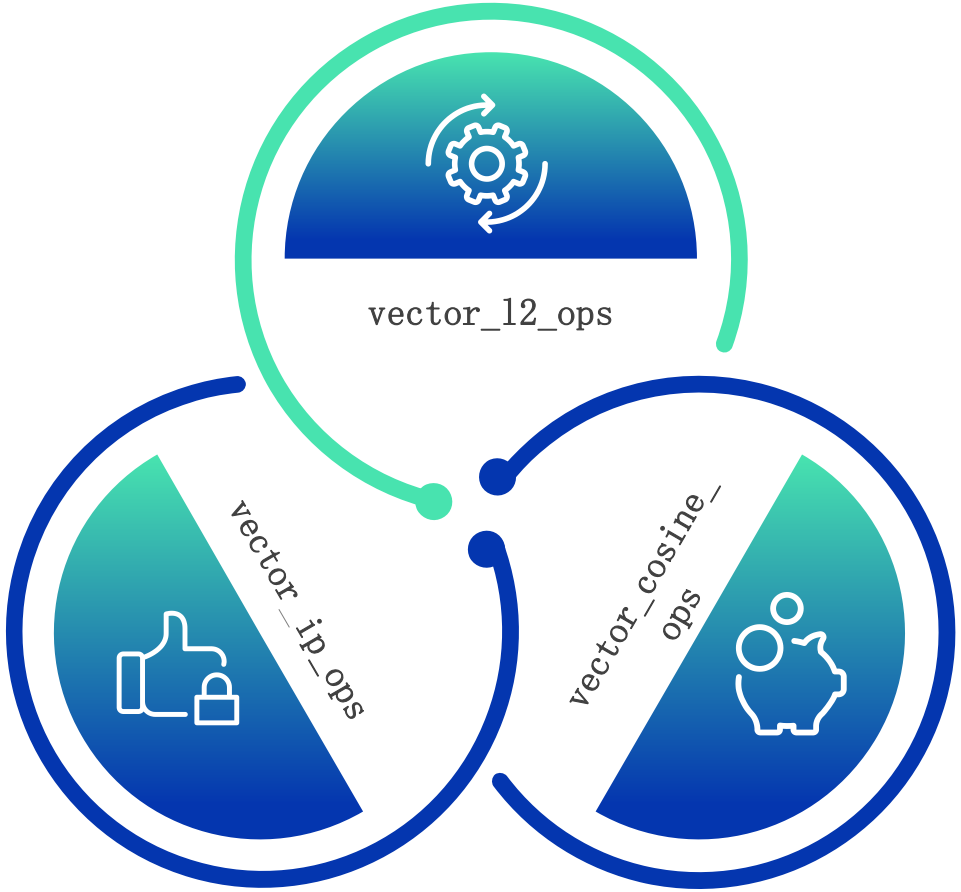
索引向量

PostgreSQL中文社区

IVFFLAT

有一个确定的数据后再建立索引，实现最佳聚类，否则性不能佳

在聚类时，向量被直接添加到各个分桶中，不做任何压缩。这种基于聚类，多簇搜索的方式搜索速度和准确性都不错。



性能最优

查询有对应的索引

向量索引是“近似最近邻”索引，SQL 必须以正确的方式在“ORDER BY”子句中使用“最近邻运算符”

索引具有最佳列表大小

建议列表大小等于 rows / 1000.

足够的内存用于索引

设置合适的 maintenance_work_mem参数

PostgreSQL中文社区

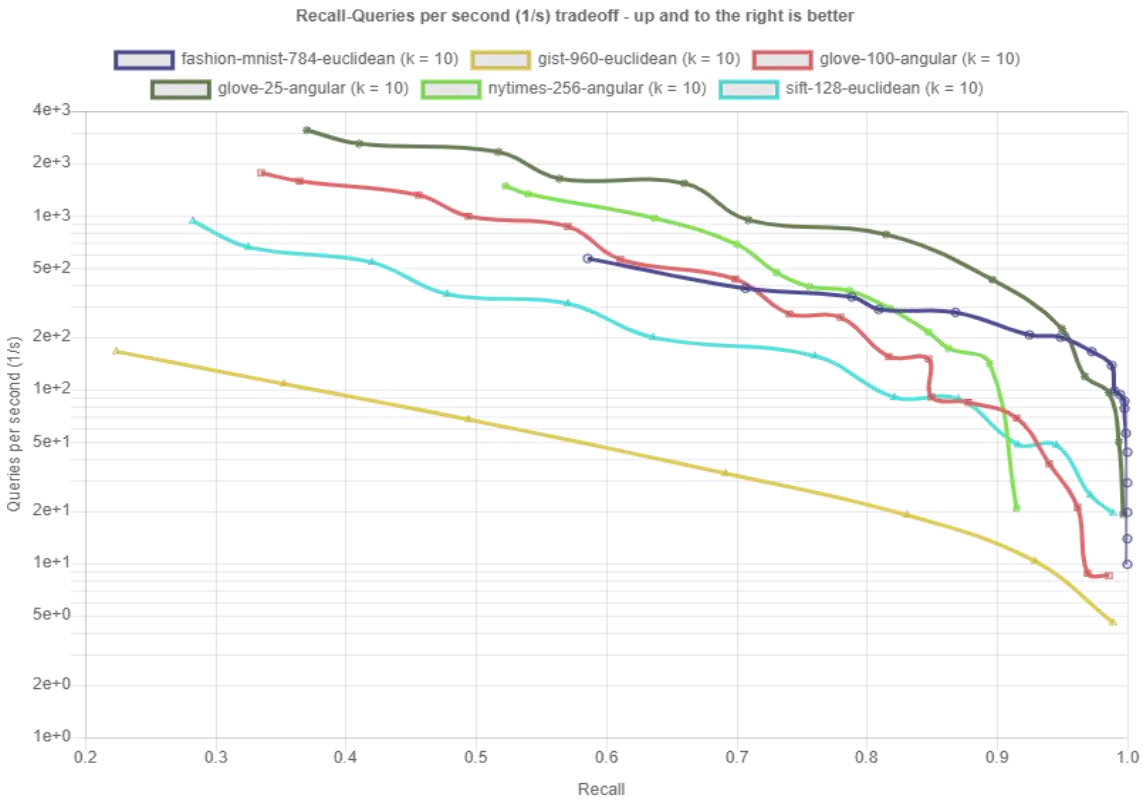
PART 03

性能测试

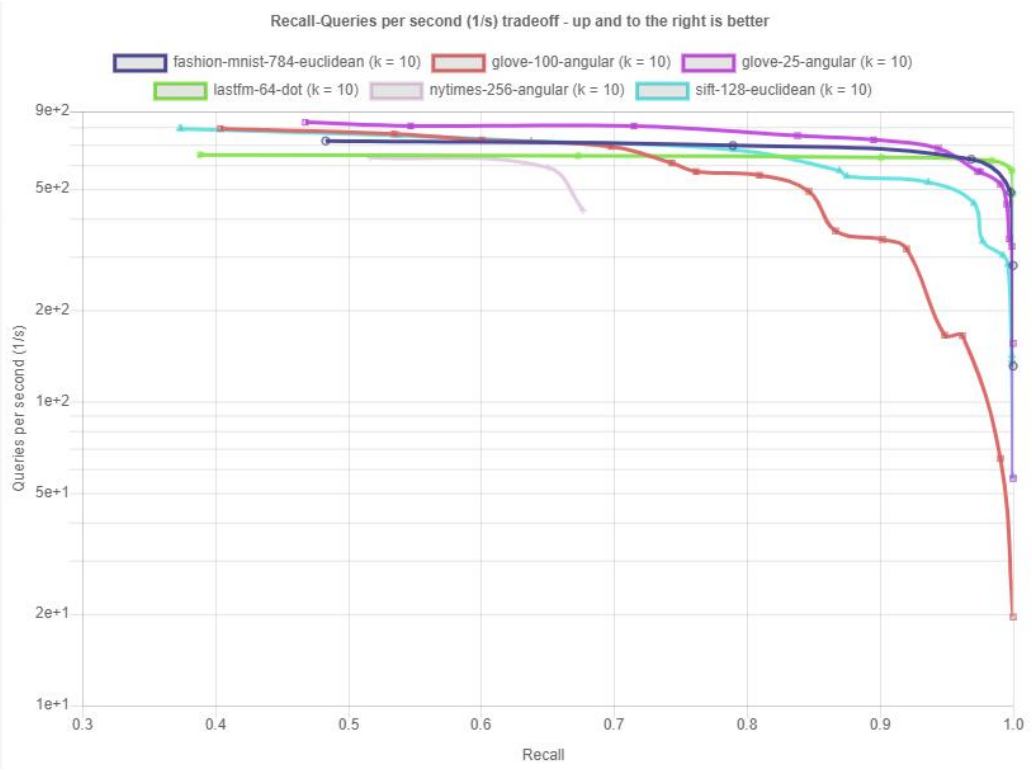
性能压测

PostgreSQL中文社区

pgvector



milvus



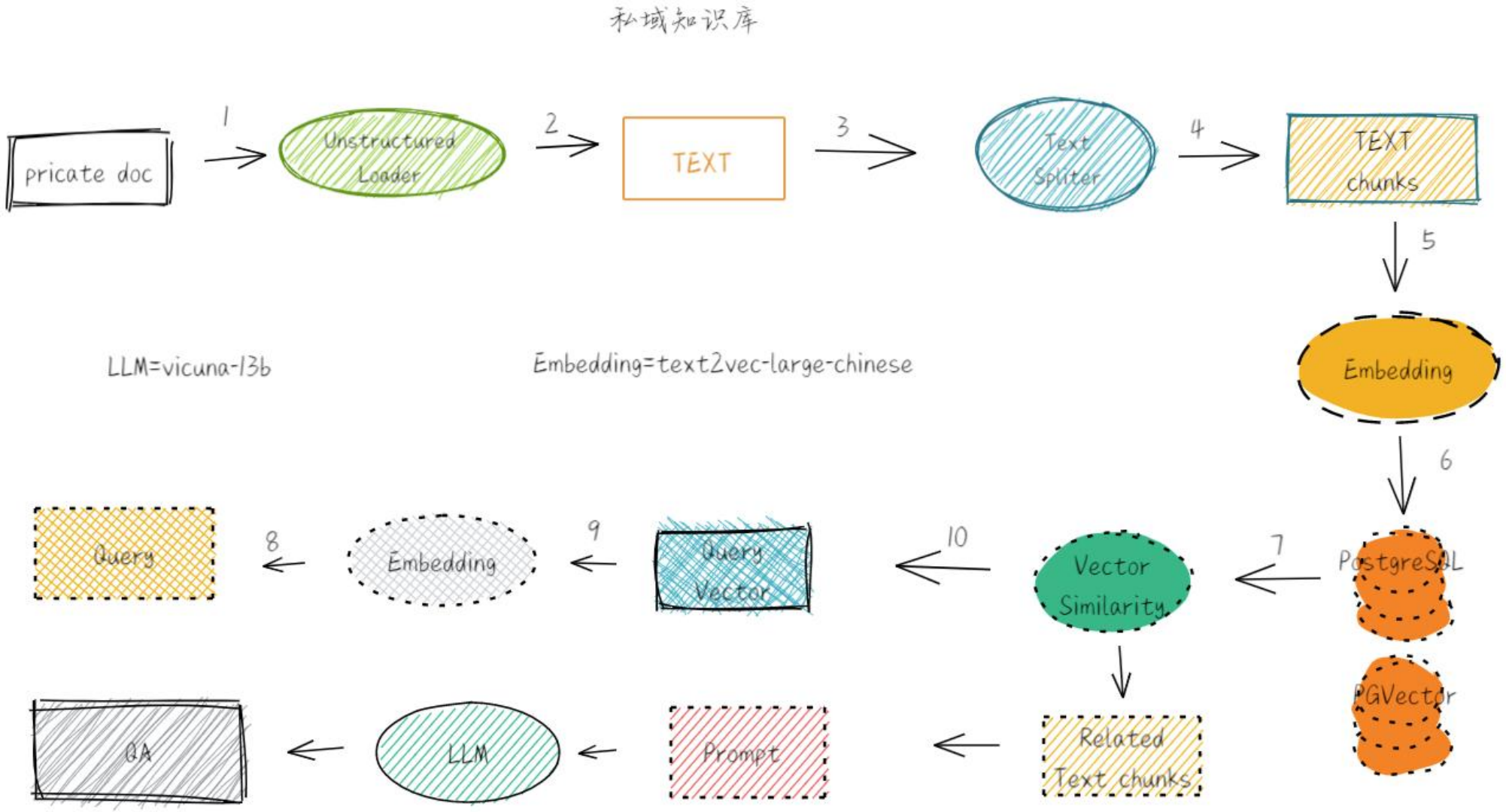
PostgreSQL中文社区

PART 04

应用案例

私域知识库

PostgreSQL中文社区



数据整合

PostgreSQL中文社区

对pdf，csv，txt，md等文档进行unstructured lodder，对文本内容进行切割，保证不超过 Embeddings 模型的 tokens 限制，拆分成多个 chunk

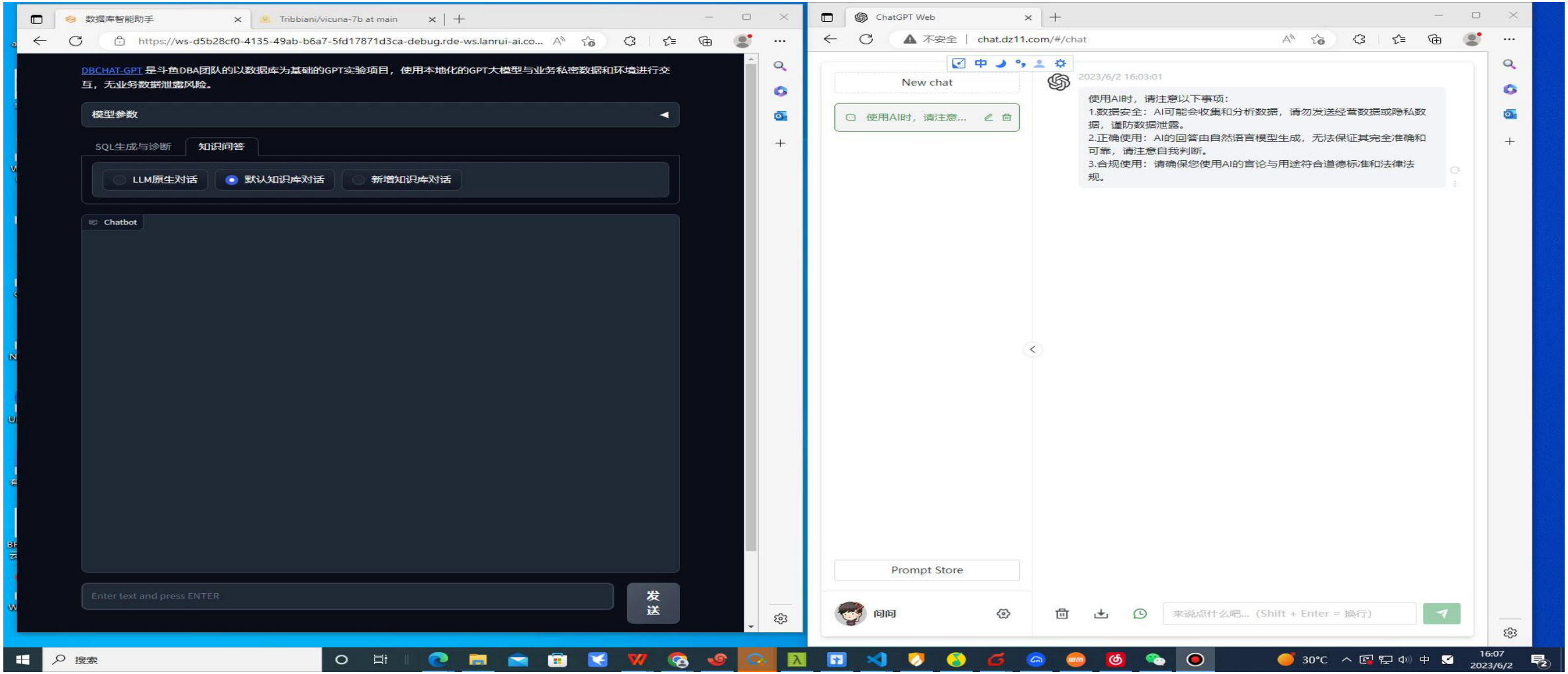
逐一对 chunk 调用 Embeddings 模型后获取对应的 vector 向量后，存储到pgvector，关键的数据主要是 vector 值即 chunk 内容

根据用户问题调用 Embeddings 模型后获取该问题的 vector，在本地向量数据库进行查询，可以获取相似度 TopN 的 chunks

把 TopN 的 chucks 作为 chat 接口的 context，再加上用户问题作为 Prompt，通过 GPT 模型获取相应的结果

案例实践

PostgreSQL中文社区



PostgreSQL中文社区

谢谢