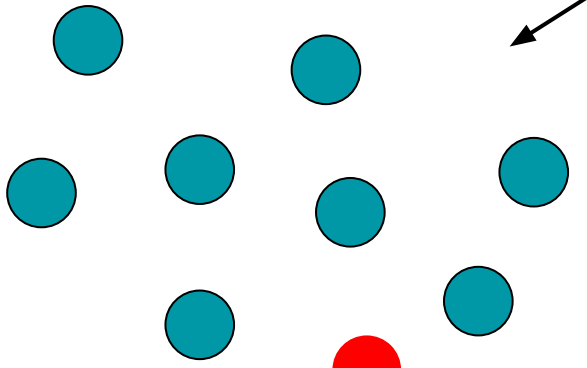
近似最近邻搜索算法研究与应用

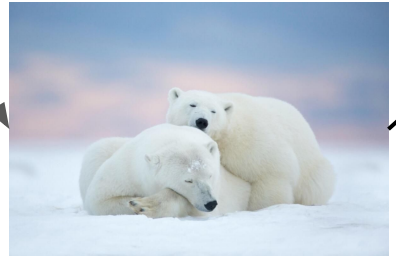
答辩学生:刘凤山 MG1833098
导师:申富饶教授

应用场景

- 搜索与推荐
- 相似特征检索



请求图片

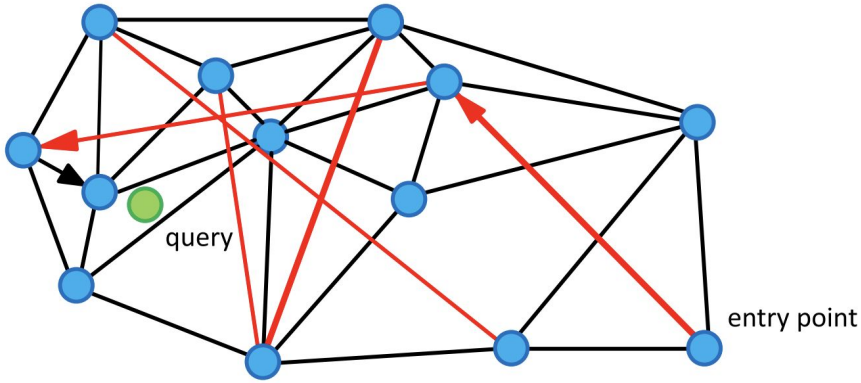


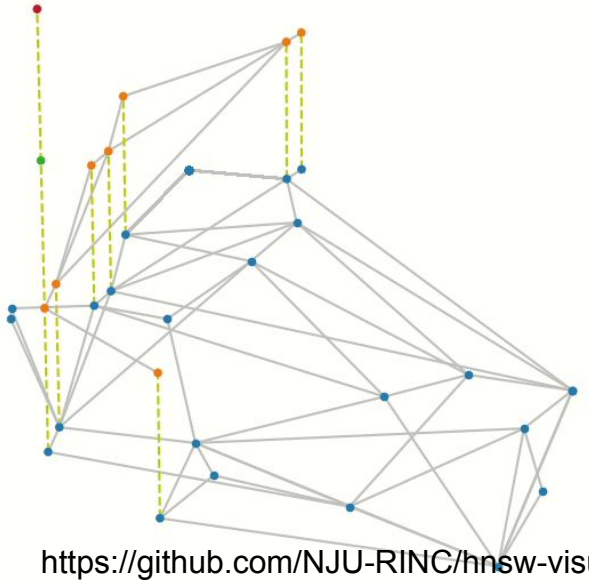
算法优化一

对HNSW的优化

HNSW

基于图的算法，目前最好的算法之一。



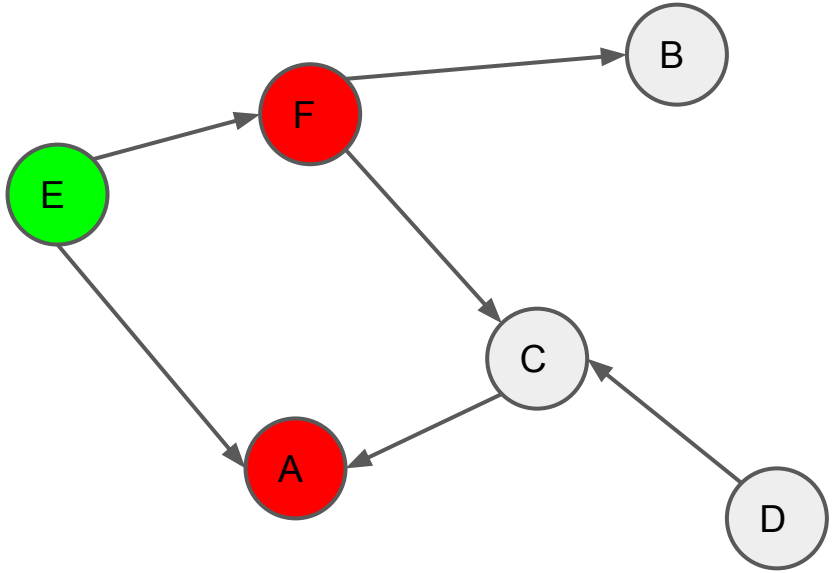


https://github.com/NJU-RINC/hnsw-visulize

HNSW存在问题：搜索返回结果不足K个

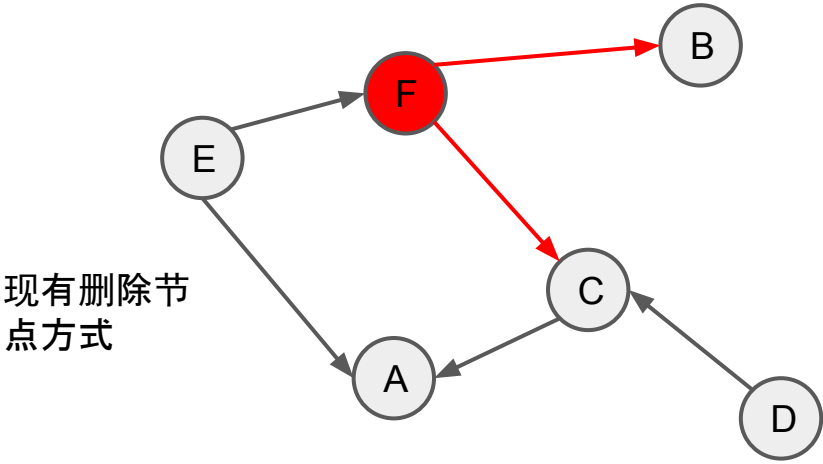
现有HNSW算法并不支持节点删除操作。节点删除只是简单的将节点标记为删除，然后搜索的时候忽略掉被删除的节点。

如右图所示，当搜索起点E的“邻居”F和A都被标记为删除之后，导致搜索直接停止，返回零个结果。

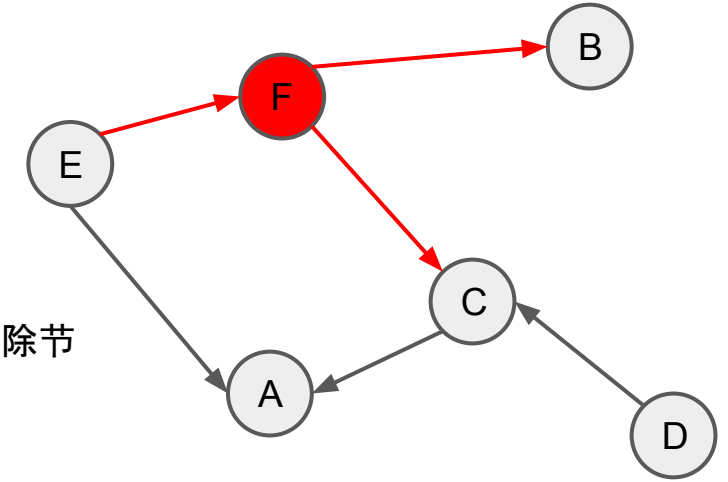


导致问题的原因

1. 根本原因
   HNSW构建的是单向图结构
2. 怎么解决
   需要知道有哪些节点拥有指向被删除节点的边



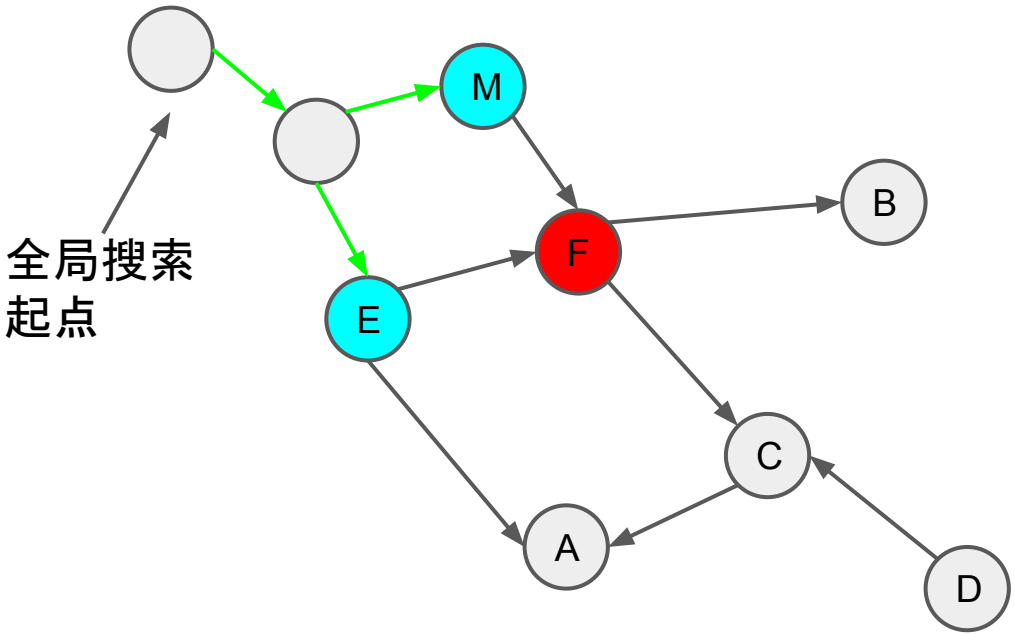
现有删除节点方式



期望删除节点方式

解决方案

1. 为了获得有哪些节点拥有指向被删除节点的边。可以将被删除节点作为搜索目标，然后在图中搜索。
2. 将被删除节点的邻居全部删除并将其坐标设置为无穷远出。
3. 对于受影响较大的节点会将其重新插入



算法效果

1. 每次请求有1000条数据，K设为10，观察返回结果数量
2. 提出的算法有效解决了该问题

数据数量与删除比例	数据维度与数据类型	HNSW 构建参数	是否使用 HNSW Mutual-Remove	返回结果不足 K 的次数
100,000 0.7	128 4 字节浮点数	$efConstruction = 200$ $m = 8$	否	44
			是	**0**
100,000 0.7	128 4 字节浮点数	$efConstruction = 200$ $m = 12$	否	0
			是	0
500,000 0.8	64 4 字节浮点数	$efConstruction = 200$ $m = 8$	否	53
			是	**0**
500,000 0.8	64 4 字节浮点数	$efConstruction = 200$ $m = 12$	否	12
			是	**0**
1,000,000 0.8	32 4 字节浮点数	$efConstruction = 200$ $m = 8$	否	61
			是	**0**
1,000,000 0.8	32 4 字节浮点数	$efConstruction = 200$ $m = 12$	否	13
			是	**0**

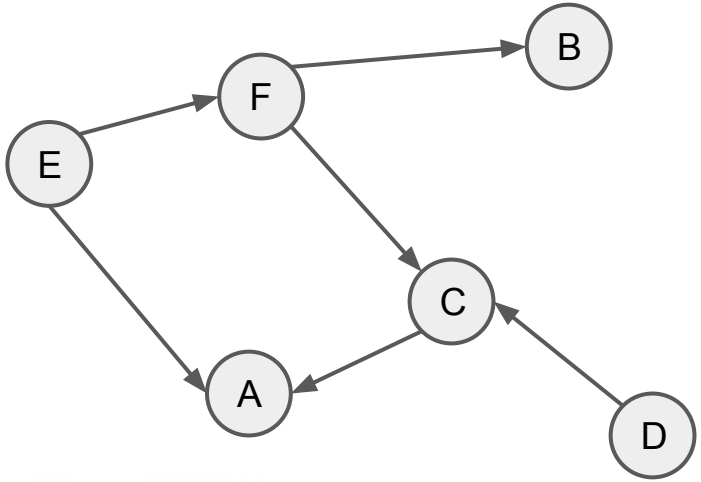
算法优化二

对IVF-HNSW的优化

HNSW存在的内存占用过大的问题

HNSW的内存开销

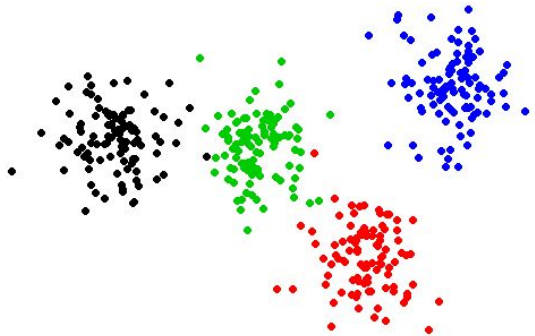
1. 高维向量
2. 图结构的存储开销



数据条数	数据维度	qps要求	内存需求	资源瓶颈
5亿	50	2000	147G	内存
3亿	50	2000	87G	内存
1.2亿	1280	200	592G	内存

非图的方法-IVF-HNSW

1. 先在真个数据集上聚类，得到nlist个聚类中心
2. 根据聚类中心，将数据集拆分成nlist份，这一过程称为数据分桶
3. 采用HNSW索引nlist个聚类中心
4. 一般数据会经过量化压缩之后再放入桶中



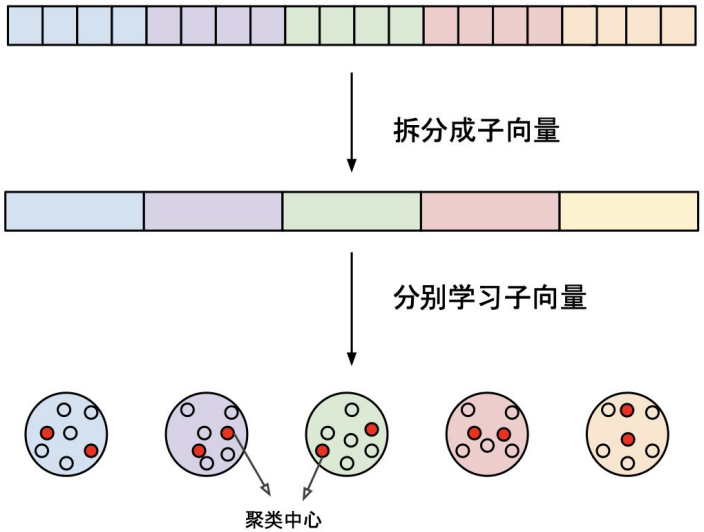
给定请求q，搜索过程如下：

1. 在HNSW索引的nlist个聚类中心中搜索距离q最近的nprobe个聚类中心
2. 依次搜索这nprobe个桶中的数据，当目前访问的数据量达到了预先定义好的上限max_codes或者全部桶搜素完成，则结束

矢量量化

对向量进行量化

1. 高维向量拆分为M个相同维度的子向量
2. 对每个子向量对应的子空间，用聚类方法学到K个代表点
3. 编码：编码的时候每个子空间里面的数据通过选取距离其最近的代表点进行表示
4. 编码后表示数据需要M*logK个比特



效果

表 4-4: HNSW_SQ、HNSW_PQ 以及 IVF-HNSW_PQ 在数据集 VIDEO-03 上的实验

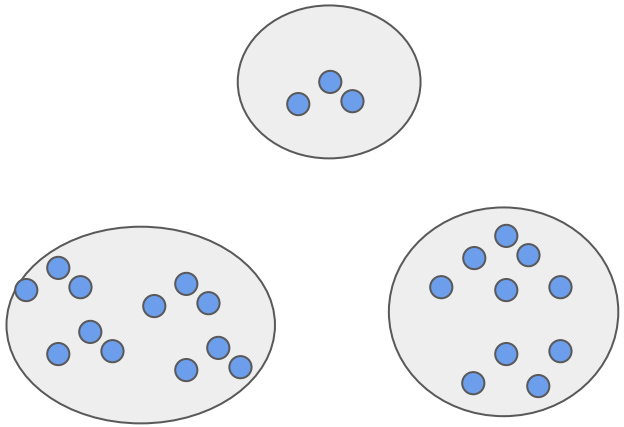
数据信息	数据名称	距离	数据维度/ 每个维度数据类型	数据数量 （万）	K	数据大小 （GiB）
	VIDEO-03	欧式距离	1280/4 字节浮点数	149	1,000	7.3

实验结果	方法	构建参数	搜索参数	训练时间 （min）	召回率 (%)	平均召回时间 （ms）	索引大小 （GiB）
	HNSW	$m = 32,$ $efConstruction = 460$	$efSearch = 1,000$	15.1	94.5	1.6	7.5
	HNSW_SQ8	$m = 32,$ $efConstruction = 460$	$efSearch = 1,000$	7.5	94.0	0.9	2.2
	HNSW_PQ320	$m = 32,$ $efConstruction = 460$	$efSearch = 1,000$	12.0	81.5	0.7	0.84
	IVF-HNSW_PQ320	$m = 32,$ $efConstruction = 460,$ $nlist = 10,000$	$nprobe = 128,$ $max\_codes = 100,000$	55.4	90.0	1.7	**0.61**

IVF-HNSW训练较慢

如何优化IVF-HNSW训练速度

- IVF-HNSW采用k-means，替换成batch k-means能够更快
- 但是替换之后会加剧数据分布的不均衡，如右图所示
- 数据分布不均衡，最终会导致系统出现较大波动，影响系统稳定性



优化算法

- 为了解决上述问题，我提出了以下解决方法
  给定数据量为n，聚类中心个数（桶个数)为m，则平均每个桶的数据条数为x_{meam}=n/m，假设低i个桶的实际数据条数为x_i，定义每个桶的不平衡系数为$d_i=x_i/x_{mean}$，预先定义一个不平衡系数的阈值Θ（大于1），预先搜索聚类中心的个数记为k
  算法如下：

  1. 使用batch-means学习m个聚类中心
  2. 对于每个数据，搜索离它最近的k个聚类中心，用从近到远的顺序遍历k个聚类中心，如果某个聚类中心的不平衡系数未达到阈值Θ，则将该数据加入该聚类中心对用的桶中

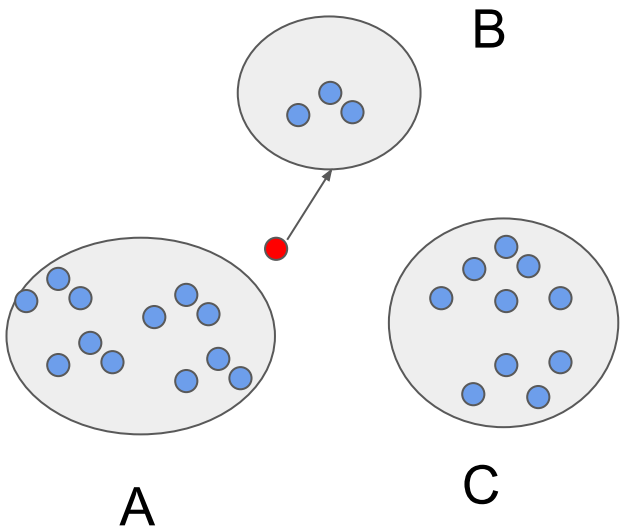


表 4-5: IVF-HNSW 与 Balanced IVF-HNSW 在数据集 TEXT-01 上的实验

数据信息	数据名称	距离	数据维度/ 每个维度数据类型	数据数量 （万）	K	数据大小 （GiB）
	TEXT-01	欧式距离	128/4 字节浮点数	1,100	1,000	5.7

实验结果	方法	构建参数	搜索参数	训练时间 （min）	召回率 (%)	平均召回时间 （ms）	索引大小 （GiB）
	IVF-HNSW_PQ16	$m = 32,$ $efConstruction = 460,$ $nlist = 262,144,$ $niter = 3$	$efSearch = 1,000,$ $nprobe = 1,000,$ $max\_codes = 10,000$	176.8	**94.7**	6.7	0.47
	Balanced IVF-HNSW_PQ16	$m = 32,$ $efConstruction = 460,$ $nlist = 262,144,$ $niter = 3, batchsize = 10,000$	$efSearch = 1,000,$ $nprobe = 1,000,$ $max\_codes = 10,000$	**36.0**	90.1	**6.3**	**0.47**

应用:微信分布式近似最近邻搜索组件SimSvr

- SimSvr = 搜索算法 + 分布式框架
- HNSW(hnswlib)，Balanced IVF-HNSW这两个算法作为核心搜索算法
- 服务现状
  - 广泛应用于微信看一看，搜一搜，视频号等业务
  - 索引总数据量超30亿，日检索量达1亿
  - Balanced IVF-HNSW今年三月已经索引9.6亿数据
  - 技术文章《微信搜索为什么这么快》阅读量过万

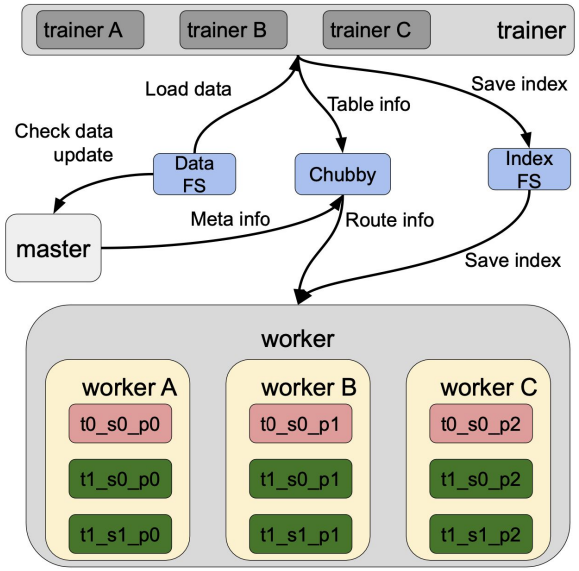


图 5-1: SimSvr 系统架构。图中箭头方向表示主要信息的流动方向。

总结

本文贡献

1. 提出HNSW Mutual-Remove，解决HNSW无法删除节点的问题
2. 提出Balanced IVF-HNSW，大幅加快IVF-HNSW构建索引的速度
3. 将1，2中提出的算法应用到了实际应用SimSvr中
4. 给HNSW的开源实现hnswlib贡献了代码

研究生期间工作

专利

1. 申富饶,**刘凤山**,赵健,李俊.“种增量式语音命令词识别方法”(201911080670.8)
2. 申富饶, **刘凤山**,邵玥. “一种基于激光雷达的液压支架对齐方法”(202010326088.1)

论文

1. An, Junyi, **Fengshan Liu**, Jian Zhao, and Furao Shen “IC Neuron: An Efficient Unit to Construct Neural Networks”, in arXiv preprint arXiv:2011.11271 (2020).
2. An, Junyi, **Fengshan Liu**, Jian Zhao, and Furao Shen “C Networks: Remodeling the Basic Unit for Convolutional Neural Networks”, in arXiv preprint arXiv:2102.03495 (2021).

开源项目

hnswlib(github 1.5kstar),contributor

谢谢！