Efficient Approximate Nearest Neighbor Search in Multi-dimensional Databases

Yun Peng
Institute of AI and BlockChain, Guangzhou University, China
Department of Computer Science, Hong Kong Baptist University, China

Byron Choi
Department of Computer Science, Hong Kong Baptist University, China

Tsz Nam Chan
Department of Computer Science, Hong Kong Baptist University, China

Jianye Yang
Cyberspace Institute of Advanced Technology, Guangzhou University, China

Jianliang Xu
Department of Computer Science, Hong Kong Baptist University, China

摘要

近似最近邻(ANN)搜索是多维数据库中的一种基本搜索，在现实世界中有许多应用，如图像检索、推荐、实体解析和序列匹配。邻近图(PG)是人工神经网络搜索中最先进的索引。然而，在已有的pg上进行搜索，要么时间复杂度高，要么对搜索结果没有性能保证。在本文中，我们提出了一个新的τ-monotonic图(τ-MG)来解决这些限制。τ-MG的新奇之处在于一个τ-monotonic属性。基于这一性质，我们证明了如果查询q与其最近邻居之间的距离小于一个常数，则在τ-MG上的搜索保证找到q的精确最近邻居，并且搜索的时间复杂度小于所有现有的基于pg的方法。为了提高索引构建效率，我们提出了τ-MG的近似变体，即τ-monotonic邻域图(τ-MNG)，它只要求每个节点的邻域为τ-monotonic。我们进一步提出了一种优化方法来减少搜索中的距离计算次数。我们广泛的实验表明，我们的技术在已知的现实世界数据集上优于所有现有的方法。

关键词：近邻图；近似最近邻搜索；τ-monotonic；边缘遮挡规则

1 引言

多维数据库中的近似最近邻(ANN)搜索是一种基础搜索，有许多应用，如图像检索、推荐、实体解析和序列匹配。已经提出了许多人工神经网络搜索方法，如基于树的方法，基于

量化的方法，基于哈希的方法，以及基于接近图的方法。最近的研究表明，在许多大规模的人工神经网络搜索应用中，基于邻近图(PG)的方法提供了优于其他方法的性能。给定一个数据库 $\mathcal{D}$，在 m-dimensional 空间中有 n 个点，基于pg的方法构造一个到索引 $\mathcal{D}$ 的接近图 $\mathcal{G}$，其中 $\mathcal{G}$ 中的每个节点对应于 $\mathcal{D}$ 中的一个点，如果两个节点满足某种接近性，则它们有一条边。对于查询点 q，可以使用 $\mathcal{G}$ 上的贪心路由来查找 q 的ANN。具体来说，在每个路由步骤中，我们计算 q 与当前节点的邻居之间的距离。然后，我们选择最接近 q 的邻居作为下一个当前节点，并继续进行下一个路由步骤。如果当前节点没有比自身更接近 q 的邻居，则路由停止。

现有的基于pg的方法主要存在以下两种极端。设 $\overline{v}$ 表示 q 最近的邻居，$\delta(q,\overline{v})$ 表示 $\overline{v}$ 到 q 的距离。在一个极端，一些研究假设 $\delta(q,\overline{v})=0$（即 q 是 $\mathcal{D}$ 中的一个点）。例如，MRNG保证通过贪婪路由找到 $\overline{v}$，期望时间复杂度为 $O\left(n^{2/m}\ln n\right)$，概率至少为 $1-(1/e)^{\frac{m}{4}(1-\frac{3}{e^2})}$。然而，假设 $\delta(q,\overline{v})=0$ 并不总是实际的，因为 $q\in\mathcal{D}$ 并不总是成立。在另一个极端，许多作品只研究设置 $\delta(q,\overline{v})<\infty$。然而，这些工作要么不能提供贪婪路由的错误保证（例如，SSG，HNSW， DPG），要么需要 $O(n)$ 时间来检索最近的邻居 $\overline{v}$（例如，DG）。

在本文中，我们研究了一个实际的设置 $\delta(q,\overline{v})<\epsilon$，它介于两个极端之间，其中 ϵ 是一个用户定义的常数。它的动机是观察到在现实世界的基准数据集中，查询通常不在 $\mathcal{D}$ 中，而是靠近它们最近的邻居 $\mathcal{D}$。例如，在我们对100万个数据点的SIFT数据集进行的初步实验中，我们观察到 $\delta(q,\overline{v})$ 与数据集中所有点的距离相比不是零，而是很小。图1(a)显示了随机选择1000个查询的 $\delta(q,\overline{v})$ 的直方图。我们可以看到对于大多数查询，$20<\delta(q,\overline{v})<270$。图1(b)为随机选择的10个查询与SIFT中所有点之间距离的箱形图。我们可以看到 q 和它最近的邻居比 $\mathcal{D}$ 上的其他点要近得多。目前唯一考虑设置 $\delta(q,\overline{v})<\epsilon$ 的是FANNG。如果 $\delta(q,\overline{v})<0.05$，FANNG 保证通过贪心路由找到 $\overline{v}$。然而，我们证明了贪心路由算法具有很高的时间复杂度。本文提出了一种求解 q 在 q 满足 $\delta(q,\overline{v})<\epsilon$ 时的最近邻 $\overline{v}$ 的时间复杂度较低的方法。

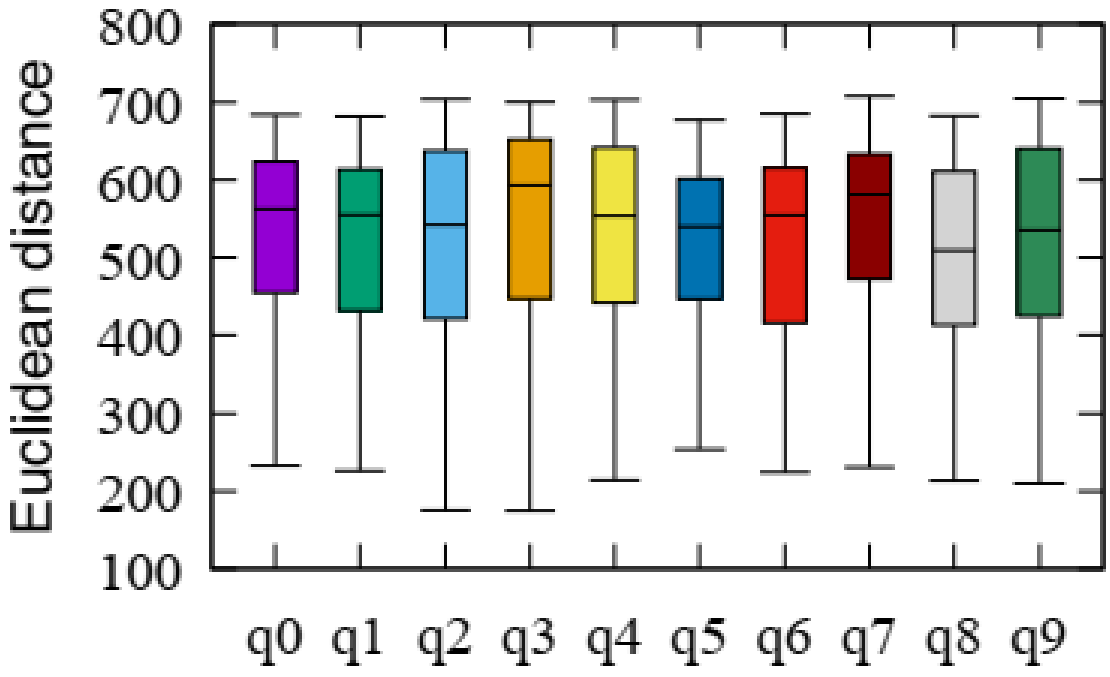

图 1. 10个查询与SIFT数据集中所有数据点之间的距离分布

2 相关工作

在本节中，我们将重点关注与邻近图(pg)密切相关的人工神经网络工作。最近对pg进行了研究。大多数现有的pg基于三种基本图模型:Delaunay图、可导航小世界图和相对邻域图。现简要回顾如下。

Delaunay图（DG）是Voronoi图的对偶图。对于 m 维欧几里得空间中的任何查询 q，DG 保证通过贪心路由找到 q 的最近邻居。然而，当 m 较大时，DG 会变成一个完全图，这使得路由很耗时。为了降低 DG 的节点度，提出了 k-nearest邻居图（kNNG）作为 DG 的近似，其中每个节点都连接到它的顶部 k 最近的邻居。例如，Jin等人和Hajebi等人分别提出了使用 kNNG 进行神经网络搜索的 IEH 和 GNNS。由于构造 kNNG 非常耗时，需要花费 $O(n^2)$ 的时间，一些研究提出构造近似的 kNNG。特别是Dong等提出了一个PG，即 KGraph 来近似 kNNG。KGraph 随机初始化每个节点的邻居，然后根据“邻居的邻居很可能是邻居”的原则，对每个节点的邻居进行迭代改进。EFANNA不是随机初始化，而是首先在数据库上构建kd树，并在kd树上使用ANN搜索来初始化每个节点的邻居，然后再执行NN-Descent。然而，KGraph 和 EFANNA 并不能保证所构建图的连通性，这会显著降低搜索结果的准确性。最近，Wen等提出了使每个节点相邻边多样化的DPG。然而，DPG 既没有降低时间复杂度，也没有为搜索结果提供错误保证。

自著名的Milgram社会实验以来，**可导航小世界图**(Navigable small world graph, NSWG)备受关注。Milgram观察到，一个大图中的两个节点通过一条短路径连接，并且可以通过贪婪路由找到该路径。人们提出了许多工作来解释和分析NSWG的性能。例如，Watts和Strogatz提出了一个二维晶格模型，并证明了晶格中两个节点之间存在一条长度为 $O(\ln n)$ 的路径。Kleinberg证

明了贪心路由无法找到路径。Kleinberg提出了另一种二维晶格模型，保证在 $O((\ln n)^2)$ 期望时间内通过贪心路由找到长度为 $O(\ln n)$ 的路径。Martel和Nguyen扩展了Kleinberg的工作以支持 m-dimensional 格。受Kleinberg模型的启发，Malkov等提出了一种PG(即NSW)来支持 m-dimensional欧几里得空间中的近似ANN搜索。但是NSW节点度高，路由开销大。NSW不保证连通性，这会影响搜索的准确性。最近，Malkov等人提出了一种分层版本的NSW(即HNSW)，以确保在多对数时间内的连通性和支持路由。然而，对时间成本的分析缺乏严谨的理论支持。此外，HNSW对搜索结果没有错误保证。

相对街区图(RNG)消除了所有可能的最长的边三角形在数据库中的点 D，也就是说，如果一个边缘 (uv) 图中，D 没有点 u 满足 $\delta(u,u) < \delta(u,v)$ 和 $\delta(u',v) < \delta(u,v)$。RNG 保证每个节点的平均程度是一个很小的常数。后来，Dearholt等证明了RNG没有足够的边来保证贪心路由搜索结果的准确性。Fu等提出了单调相对邻域图(monotonic relative neighborhood graph, MRNG)。MRNG 保证每个节点的平均度是一个常数，并且保证在 $q \in D$ 时找到 q 最近的邻居。然而，当 $q \in D$ 时，MRNG 对搜索结果没有错误保证。FANNG确保在 $\delta(q,\overline{v}) < \tau$ 时找到 q 的最近邻居 $\overline{v}$。然而，FANNG 并没有对贪婪路由的节点度和时间复杂度进行理论分析。最近，Fu等将MRNG扩展到卫星系统图(SSG)。虽然SSG是针对任意 $q \notin D$ 设计的，但对于贪心路由的搜索结果，SSG 没有错误保证。

一些研究改进了PG上的路由算法，例如Muñoz等提出在每个路由步骤中修剪与q不在同一象限的邻居。Baranchuk等使用图神经网络选择路由到的邻居。Peng等人使用神经网络来修剪没有希望的邻居，以减少距离计算的次数。Li等提出了一种基于学习的提前停止路由的方法，以避免不必要的路由步骤。但是，这些工作对搜索结果没有错误保证。Zhao等和Yu等提出使用GPU来加速PG上的路由，但本文主要关注的是CPU，与他们是正交的。

Non-PG-based方法。也有一些不基于PG的人工神经网络作品，如基于树的方法，基于倒排索引的方法，基于量化的方法，以及基于哈希的方法。由于最近的研究表明这些方法优于基于PG的方法，因此我们不将这些方法包括在本节中。我们建议感兴趣的读者参考优秀的调查以了解更多细节。

已有关于人工神经网络搜索意义的研究。随着维数的增加，对比(即查询q与其最近点和最远点之间的距离之比)趋于1,ANN搜索变得没有意义，因为q的NN无法与其他点分离。然而，如果数据集的内在维数较低，或者查询与其最近邻居之间的距离不大于一个常数，则存在对比，ANN搜索是有意义的。

3 本文方法

定义 1. (边缘遮挡规则 MRNG) 给定图 G 的三个节点 u, u' 和 v，如果 $(u,u') \in G$ 且 $u \in \text{ne}(u,v)$，则 $(u,v) \notin G$。另外，我们说 (u,u') 遮挡 (u,v)。

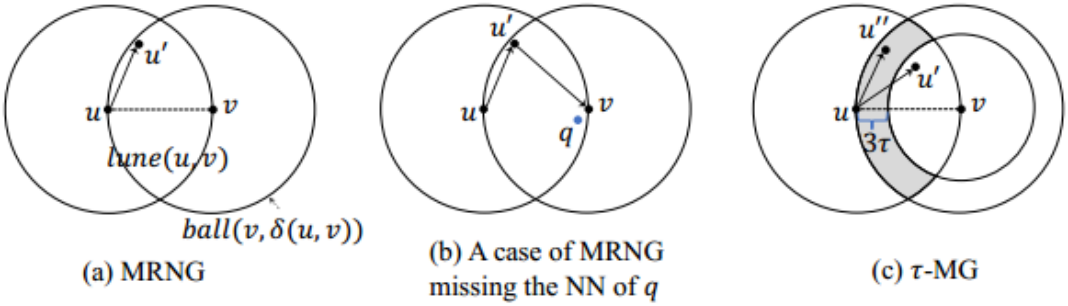

图 2. MRNG和 τ-MG 的边缘遮挡规则说明。在 (a) 和 (c)，边缘 (u, u') 可以封闭边缘 (uv)。在 (c)，边缘 (u, u') 不能遮挡 (u, v)。(b) 显示了贪婪搜索由于 MRNG 的边缘遮挡规则而错过了 q 神经网络的情况。

定义 2. (MRNG) 给定一个数据库 D，邻近图 G 是一个 MRNG。对于任意的两个点 $u, v \in G$，如果 G 有一条边 (u, u') 遮挡 (u, v)，则满足 $(u, v) \notin G$。

MRNG的性能保证如下。

引理 1. 如果 $q \in D$，则 MRNG 上的贪婪搜索从任意节点开始查找 q。然而，若 $q \notin D$，则在 MRNG 上进行贪婪搜索可能找不到 q 的最近邻居。

图 3(b) 给出了一个在 MRNG 上贪婪搜索无法找到 q 的 NN 的例子。边缘 (u, v) 被 (u, u') 遮挡。假设 u 是贪婪路由的当前节点。因为 $\delta(q, u) < \delta(q, u')$，贪婪路由将停止并返回 u。然而，v 是 q 的 NN。

我们注意到，尽管 MRNG 是为欧几里得空间设计的，但 MRNG 仍然可以用于一般度量空间。然而，对时间和空间复杂性的 MRNG 分析并不适用于一般度量空间。

4 复现细节

4.1 与已有开源代码对比

本文未开源，不过本文的工作是建立在NSG算法基础上的。

NSG源码：https://github.com/ZJULearning/nsg

在本文中需要实现的工作有三个，分别如下。

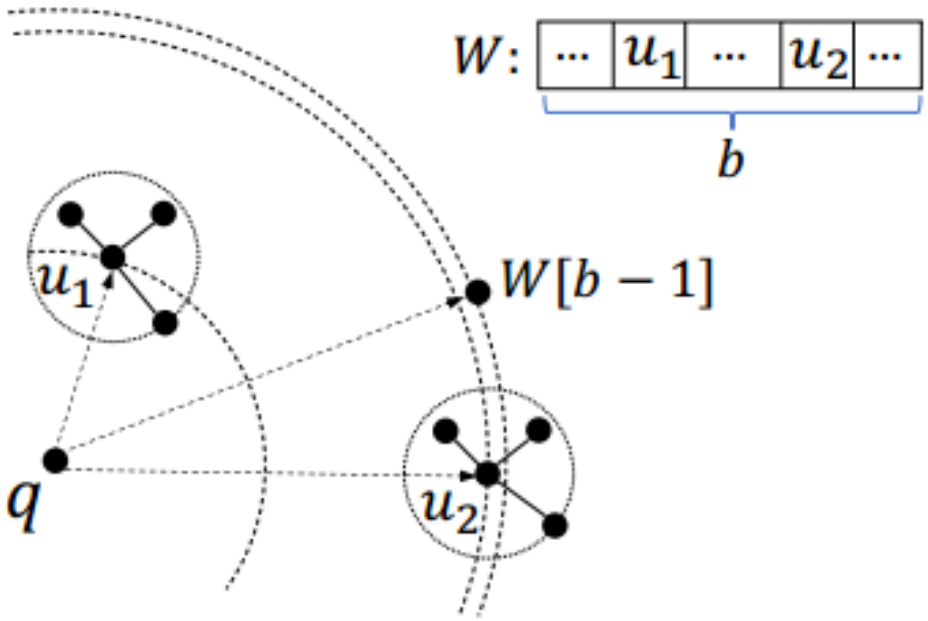

图 3. 查询感知边缘遮挡示意图 (u_1 和 u_2 是 W 中的两个节点；$W[b-1]$ 为 W 中的最后一个节点；实线表示 PG 中的边)

查询感知边缘遮挡方法(QEO)在实验中观察到当前节点 u 离 q 越远，u 的邻居被 W 中的第 $(b-1)$ 个节点修剪的概率就越高。如图二所示，也就是说，所有的邻居 u_1 不能通过 $W[b-1]$被修剪。u_2 远非 q 和 $ball(u_2, \delta(q, W[b-1]))$ 部分在 $ball(q, \delta(q, W[b-1]))$，也就是说，一些$u_2$邻居可以被$W[b-1]$修剪。

基于部分距离的修剪（PDP）。在定向搜索中，当前节点的邻居 v 只要我们可以确定 $\delta(q,v) > \delta(q, W[b-1])$，我们可以修剪 v 而不需要准确计算 $\delta(q,v)$。这可以节省大量的计算成本，并激发了基于部分距离的剪枝方法。具体来说，在 m 迭代中计算总和 $\sum_{i=0}^{m}(v[i]-q[i])^2$，如果我们发现 $\sum_{i=0}^{j}(v[i]-q[i])^2$ 在第 j 次迭代已经大于 $(\delta(q, W[b-1]))^2$，我们可以简单地删除 v。

前缀内积索引（PII）。计算 $\delta(qv)$ 可以等价于 $\langle q,q\rangle + \langle v,v\rangle - 2\times\sum_{i=0}^{m}(v[i]\times q[i])$，其中 $\langle\cdot,\cdot\rangle$ 表示内积。$\langle v,v\rangle$ 可以离线计算。

我们只需要计算 $\langle q,q\rangle$ 和 $v[i]\times q[i]$ 在线。由于计算 $v[]\times q[]$ 只需要计算 $(v[i]-q[i])^2$ 的一半操作，并且计算 $\langle q,q\rangle$ 的成本可以由搜索中的所有距离计算共享，因此可以节省大约一半的距离计算总成本。为了与基于部分距离的剪枝相结合，我们将向量 v 划分为若干段，并对前缀段的内积进行索引。具体来说，对于一段尺寸参数 s，我们计算内积 $\langle v[0, i\times s], v[0, i\times s]\rangle$，其中 $0 < i < \lceil m/s \rceil$。我们一段一段地执行基于部分距离的剪枝。

4.2 实验环境搭建

由于本论文代码是基于NSG源码的，故只需要按照NSG的安装步骤。首先我们需要建立一张KGraph图，然后需要先运行一个建立索引的程序，将KGrapg图进行修剪，然后再运行另

一个程序进行查询。实验在2核CPU，2G RAM的ubuntu22.04上运行。召回率越高，但查询延迟越高。我们在实验中关注 $K = 100$。实验的数据集为SIFT，128维的1M个向量点。

5 实验结果分析

当评估图搜索算法的性能时，召回率（Recall）和每秒查询数（Queries per Second）是两个重要的性能指标。

召回率（Recall）：

定义：召回率是指在所有实际相关的数据中，算法成功找到并返回的比例。它衡量了算法在找到所有相关结果方面的能力。在图搜索算法中，通常表示为正例（相关结果）被正确找到的比例。计算公式：召回率 = 正确找到的相关结果数量 / 所有实际相关的结果数量。

每秒查询数（Queries per Second）：

定义：每秒查询数是指算法在单位时间内能够处理的查询数量。这个指标反映了算法的查询处理速度。计算公式：每秒查询数 = 总查询数量 / 算法执行的总时间。在图搜索领域，尤其是在大规模数据集和实时应用中，高召回率和高每秒查询数都是关键的性能指标。高召回率确保算法不会错过重要的信息，而高每秒查询数则表明算法在实时应用中的效率。通常，这两个指标之间存在权衡关系，优化其中一个可能会对另一个产生影响。

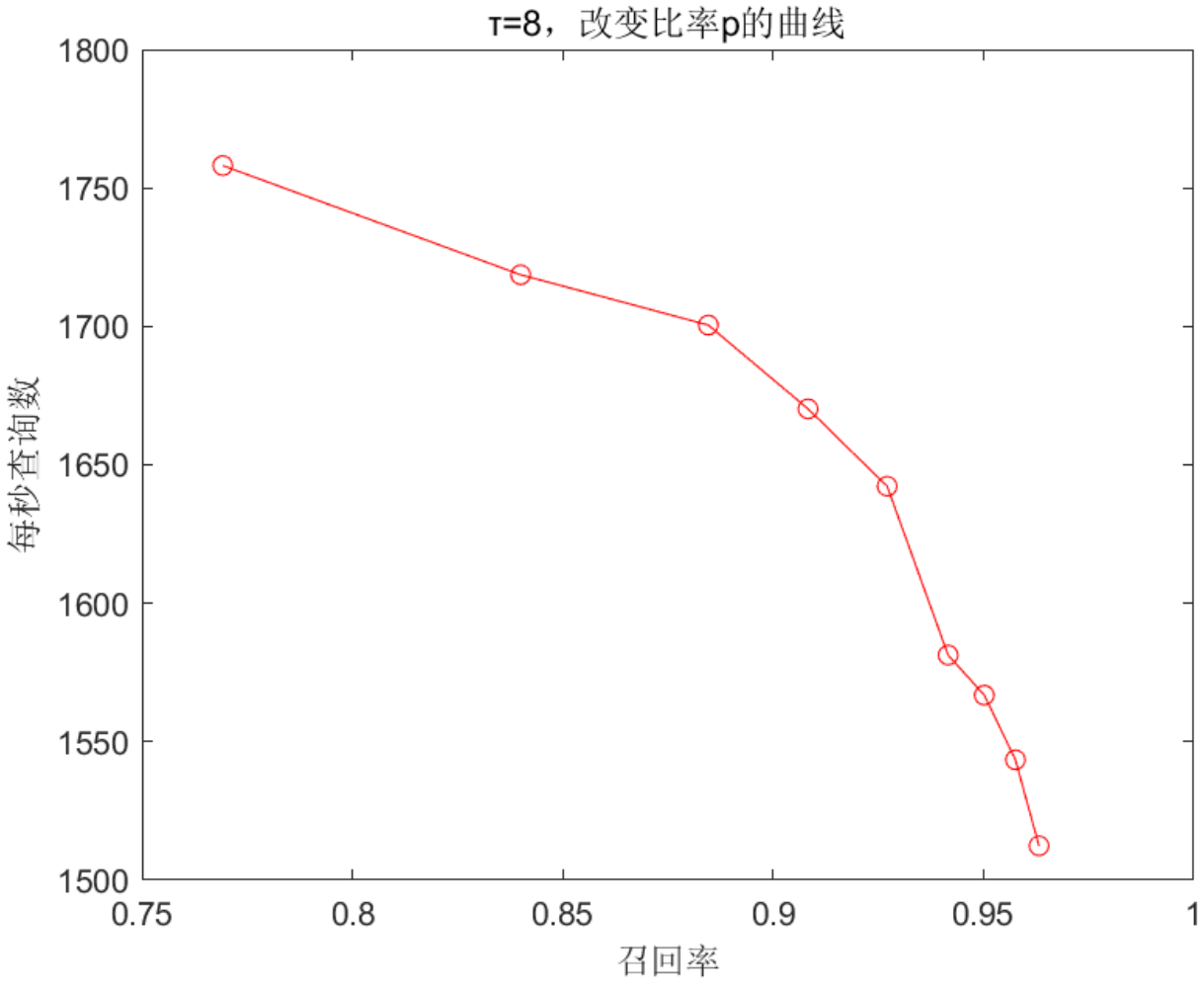

图 4. 该实验是在τ=8的情形下进行的，从左到右的数据点的概率p依次为0.1-0.9

从图3可以看出，随着p的增大，召回率也在逐渐增大，原因是p增大时，就不会有太多的节点没有计算距离而直接被修剪，但由于p增大，越来越多的节点需要距离计算，此时每秒查询数也会随之降低。

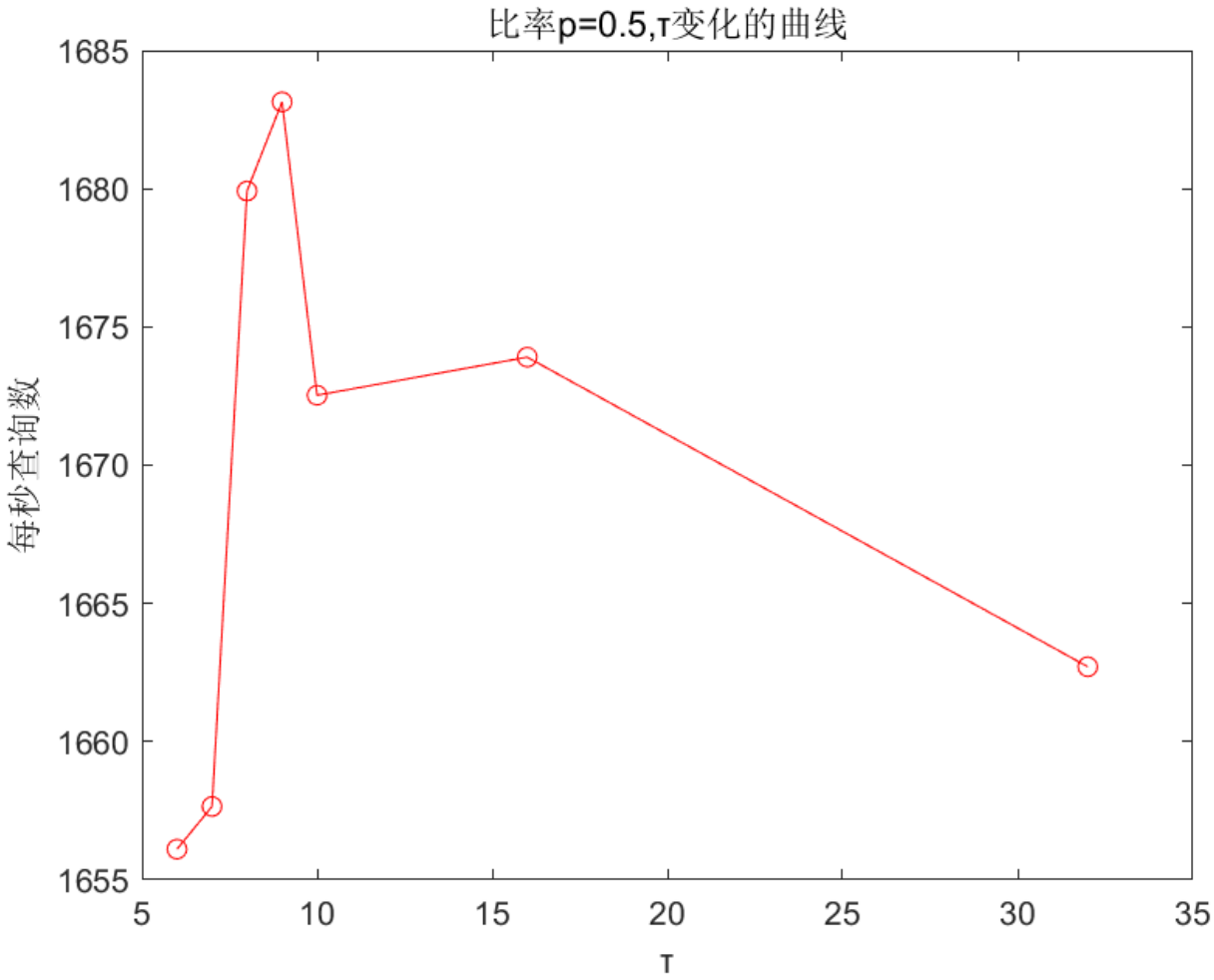

图 5. 该实验是在p=0.5的情形下进行的，从左到右的数据点的τ依次为6，7，8，9，10，16，32

从上图可以看出，τ-MNG的性能随着数据的增长先提高后下降。其原因是搜索成本由搜索中的距离计算次数(NDC)主导，期望NDC由期望搜索步数与PG的期望节点度的乘积所约束，如果增大了，则τ-MNG的节点度增大，τ-MNG的连通性更好。搜索的弯路更少，减少了NDC。但是，如果进一步增大，节点度会变得太大，搜索每一步都要计算更多邻居的距离，这将导致一个很大的NDC。

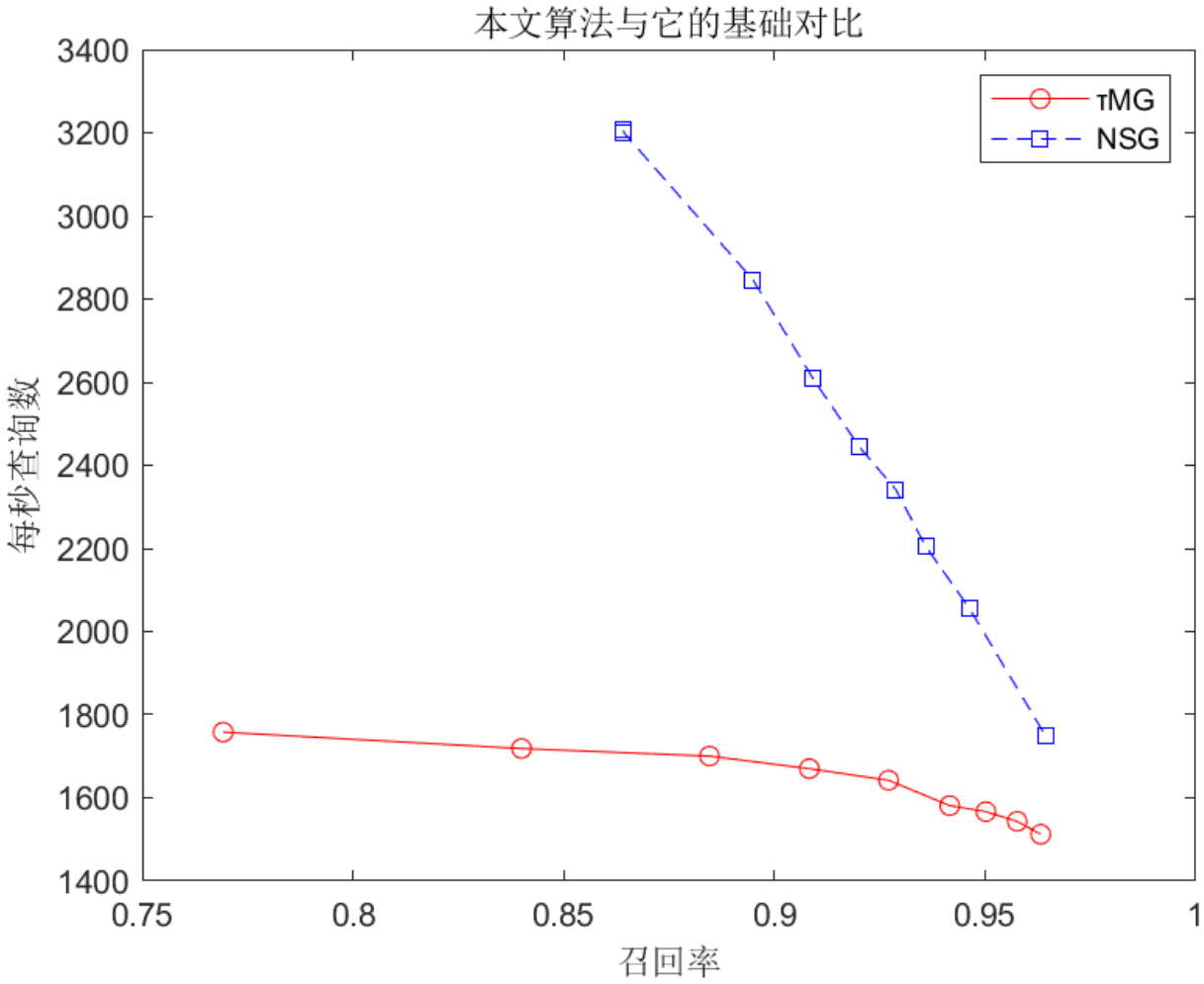

图 6. 本文的方法和本文的基本方法对比

可以看到本文的方法$\tau - mg$和基础的方法NSG速度差距很大，最大的时候接近一倍的速度差，原因是该论文的方法和基础方法其实没有什么显著的差距，在索引方面，仅仅是添加了一个τ用于边的修剪，再无其他优化。其次优化都放在了查询过程中，但本文提到的三个优化方法并没有起到很大的作用，在同样的召回率情况下，$\tau - mg$速度远不如基本方法。

6 总结与展望

在本文中，我们提出了一个 τ-monotonic 图（τ-MG）用于多维数据库中的近似最近邻搜索。τ-MG 的核心是一种新的边缘遮挡规则。当 q 到数据库中最近邻居的距离小于一个常数时，在 τ-MG 上的贪心路由保证找到 q 的精确最近邻居，并且期望的搜索时间复杂度小于所有现有的方法。对 τ-MG 中贪心路由的期望长度和 τ-MG 的期望节点度进行了严格的分析，并在补充资料中给出。为了提高索引构建的效率，我们提出了一个 τ-monotonic 邻域图（τ-MNG），它是 τ-MG 的近似变体。我们进一步提出了一种优化方法，以减少在 τ-MNG 上搜索的距离计算次数。我们的大量实验表明，我们的方法是有效的，并且在现实世界的基准数据集上优于最先进的人工神经网络搜索方法。