# 1 引言

目前，形如集合，序列，树和图等结构化数据给像高效存储，索引，部分查询 (如子图/超图搜索) 和相似性搜索之类的传统数据管理领域提出了一个巨大的挑战。随着图数据库的迅速发展，在图数据库中的**图相似性搜索**成为了一个日益重要的研究课题。图相似性搜索已经多个领域进行了应用，例如化学，分子信息学，传感器网络管理，社交网络管理，XML 文档等等。在化学与制药领域，每天会产生大量的化学分子结构数据。一旦一个新的化学结构被合成后，这个结构的特性就可能通过查询已知的分子结构，通过其特性来预测。大规模图数据中的相似性搜索让科研人员和工程师们可以对图精确建模，认识到图数据间的固有联系，减少大数据库的计算代价。

图查询需要分成两种：(i) 子图查询 (ii) 相似度查询。**子图查询**目的是确定一个包含着查询图的集合。**相似性查询**是根据差异距离来找出数据库中和查询相似的图，如 $k$-NNs 查询 (最相近 $k$ 查询) 和范围查询。$k$-NNs 查询是为了得到最相近的 $k$ 个图。范围查询，是为了得到差异距离小于预设值的所有图。在本文中，我们只研究在大规模图数据库中的 $k$-NNs 相似性搜索。

在图上进行相似性搜索是十分富有挑战的。我们认为一个理想相似性搜索设计应该达到以下三个相关的 (有时是相反的) 目标: 准确度高，时间短，空间小。为了准确度高，我们要求相似性度量方法应该抓住数据的固有相似度。众所周知，一些图上的操作，像子图同构，都是 NP 问题。所以为了时间短，我们需要设计一个高效的算法来尽量避免全图搜索。为了空间小，所以索引结构不能对图数据库开销太大。

许多简单的图相似性度量方法是将图转换为在高维欧几里得空间 (**特征空间**) 中的表示，然后利用空间上的索引技术来进行相似度搜索。现有的特征提取方法大多数都可以归为两类:(i) 每幅图的子结构单独计数（如从一幅图中生成随机的通路），(ii) 图集合中的子结构一起计数（如挖掘频繁子结构）。尽管相似性搜索已经被广泛应用了，但是在特征提取的适配性和特征的索引策略上仍有很多限制。首先，图数据库（尤其是大规模的图数据库）的特征提取是需要大量运算量的。其次，特征提取过程可能产生很多特征值，需要占用大量内存空间。学术界曾提出过许多不同的特征选取方法来确定“有辨识度的”特征。但是，对于数据库搜索表现优异的特征选取方法在图相似性上不一定会很好，所以我们还需要自行去判断与权衡。

在本文中，我们探索了一种新的图相似性搜索方法。我们通过核函数来定义相似度。和通常不同的是，**核函数**没有直接提取特征值，而是将数据映射到一个高维的函数空间，并利用在这个函数空间中数据的内积来计算其相似度。

基于核函数的图相似度测量方法的优势在于核函数统计学表现非常好，例如可以高精度的分类。但是将核函数用于数据库搜索也有些问题，主要的难点在于（i）核函数用在图上计算量很大（ii）目前为止，没有一个明确的方法来标引大规模图数据库上的核函数计算。

我们的方法称作 G-hash。我们致力于提出一种新的核函数来高效计算大规模图数据库上的特征。在我们的模型中，图被用核函数直接压缩成一个节点集。传统上，对于复杂的图结构，这样的压缩方法会丢失大量的信息。但是我们将大多数拓扑信息压缩到了每个节点的特征向量上来避免大量信息的丢失。这种方法提供了一个对于信息量丰富的图的简单表示，并且也很好进行比较。我们接着利用压缩集表示法将图节点哈希化。哈希的键值是基于这些表示集的，所以相似的节点会被哈希到同一个格子或者相邻的格子中。一旦我们将数据库中所有图哈希成一个表后，我们可以找到所有相似的节点，然后利用它们计算查询图和数据库中图的距离，并找出查询图的 $k$-NNs(k 个相近图)。

总体而言，我们这篇文章的主要贡献有:

- 提出了一个图核函数和用来进行快速图相似性搜索的索引结构
- 我们的索引只需线性时间去计算（对数据库中的总结点数而言），并且可以通过动态增删来在线构建。
- 我们证明了新的图核函数和相应的索引结构可以更好地在抓取固有图相似性和对于大数据库的快速计算上平衡。

本文组织如下，我们首先在第二章回顾了一些相关领域，如哈希，索引，图核函数等等。在第三章，我们正式定义了图和图相似性搜索。在第四章我们讨论了索引结构和核函数的细节。最后我们对我们的算法进行了广泛的实验，并与现有算法进行对比，并得出了一些结论来指导我们的研究。