**Abstract**

集合，列表，树和图之类的结构化数据给数据管理这一基础领域提出了一个严峻的挑战。包括有效存储，索引，相似性搜索等都面临着一系列问题。随着图数据库的快速增多，图相似性搜索成为了一个热门的话题。图相似性搜索在多个领域都有应用，像化学结构，分子结构，传感器网络，XML 文档分析等等。

许多现在的图索引方法都是基于子图结构的，即找到一组包含查询图的图。因此，我们不能直接利用其进行相似性搜索。在数据挖掘和机器学习领域，大家提出了很多图核函数来判断图之间的固有相似度。尽管图核函数在监督学习领域用于准确预测和分类模型效果很好，但是如果用于图相似度查询则会存在两个关键问题:（i）非常高的计算复杂度（ii）在图索引上的复杂度也是非平凡的。

我们打算构建一个图核函数和图相似性搜索的桥梁。我们打算提出两个关键解决方案（i）一种新颖的图相似度度量方法（ii）一个图数据的有效索引策略。我们的相似性度量方法是居于每个节点和其邻接节点的特征的，并且我们利用哈希表来进行高效存储和快速查询。利用图核函数抓住图相似性的本质特征，利用哈希表加速查询过程。我们将我们提出的这种方法称为 G-Hash，并且已经在大规模的化学结构数据库上进行了测试。结果表明 G-Hash 在 k 个最相近邻居问题上已经达到了业界的最高水平，更重要的是，我们这种新的相似性度量方法和索引结构比现有的算法（C-tree，gIndex，GraphGrep 等）具有更小的索引尺寸，更快的查询速度。

**关键词: 图相似性搜索，图分类，哈希，图核，k-NNs 查询**