

1 基于哈希的快速图相似性搜索

正如之前所说，现在的图查询算法查询时间很快但是没有好的相似性度量方法。核函数可以提供一个好的相似性度量方法，但是核函数的矩阵运算需要大量的时间，所以我们很难直接利用其来建立索引。为了解决这个问题，我们提出了一种新的算法，G-Hash。现在的算法常常关注速度或精度的一种，而G-Hash 在两者效果均很好。我们利用小波图匹配核 (WA) 来定义相似性，并用 Hash 表作为索引结构来加速图相似性查询。下面我们先介绍下 WA 方法。

1.1 小波图匹配核简介

WA 方法的思想是先通过压缩每个节点周围邻接节点的属性信息，然后应用非递归线性核去计算图之间的相似度。这个方法包含两个重要的概念： h -hop 邻域和离散小波变换。用 $N_h(v)$ 标记一个节点 v 的 h 跳邻域，代表一个距离节点 v 最短距离是 h 跳 (跳过 h 个节点, 也就是 $h+1$ 的曼哈顿距离) 的节点集合。离散小波变换涉及到一个小波函数的定义，见下文公式(1)，也用到了 h 跳邻域。

$$\psi_{j,k} = \frac{1}{h+1} \int_{j/(k+1)}^{(j+1)/(k+1)} \varphi(x) dx \quad (1)$$

$\varphi(x)$ 代表 Haar 或者 Mexican Hat 小波函数， h 是在将 $\varphi(x)$ 在 $[0,1)$ 区间上分成 $h+1$ 个间隔后的第 h 个间隔， $j \in [0, h]$ 。基于以上两个定义，我们现在可以将小波分析用在图上了。我们用小波函数来计算每个节点的局部拓扑和。公式(2)展示了一个小波度量方法，记做 $\Gamma_h(v)$ ，以图 G 中的一个节点 v 为例。

$$\Gamma_h(v) = C_{h,v} \times \sum_{j=0}^k \psi_{j,k} \times \bar{f}_j(v) \quad (2)$$

其中, $C_{h,v}$ 是一个归一化因子

$$C_{h,v} = \left(\sum_{j=0}^h \frac{\psi_{j,h}^2}{|N_h(v)|} \right)^{-1/2}, \quad (3)$$

$\bar{f}_j(v)$ 是离节点 v 最远距离为 j 的原子特征向量的平均值

$$\bar{f}_j(v) = \frac{1}{|N_j(v)|} \sum_{u \in N_j(v)} f_u \quad (4)$$

f_u 表示节点 v 的特征向量值。这样的特征向量值只会是下面四种中的一种：定类，定序，定距，定比。对于定比和定距特征值，直接在上述的小波分析时代入

其值即可得到局部特征值。对于定类和定序节点特征，我们首先建立一个直方图，然后用小波分析提取出特征值。在节点 v 分析完成后，我们可以得到一个节点列表 $\Gamma^h(v) = \{\Gamma_1(v), \Gamma_2(v), \dots, \Gamma_h(v)\}$ ，我们称其为小波测度矩阵。用此方法我们可以将一个图转换为一个节点向量集合。因为小波变换有明确的正负区域，所以这些小波压缩特征可以表示出局部的邻接节点和距离较远的邻接节点的差异。因此，通过小波变换，一幅图的结构化信息可以压缩成节点特征。从而我们可以忽略拓扑结构来专心于节点匹配。核函数就是建立在这些集合上的，我们以图 G 和 G' 为例，图匹配核函数是这样的

$$k_m(G, G') = \sum_{(u,v) \in V(G) \times V(G')} K(\Gamma^h(u), \Gamma^h(v)), \quad (5)$$

$$K(X, Y) = e^{\frac{-\|X-Y\|_2^2}{2}}. \quad (6)$$

就像在后文实验部分我们验证的一样，WA 方法展示了一种很好的利用核函数进行图相似度定义的方法。但是这个方法有一个问题，就是小波匹配核的总时间复杂度是 $O(m^2)$ ，核矩阵的是 $O(n^2 \times m^2)$ ， n 是数据库图个数， m 是平均节点个数。这意味着，当数据库尺寸增加时，计算时间将大幅度增加。