# 1 相关研究

在本章中，我们讨论一些相关的研究。从基本的图数据库索引，子图搜索，模糊子图搜索和图相似性搜索到图核函数

## 1.1 子图搜索

现在许多的子图搜索都采用了一个相似的框架，先把图分解为一系列小片段，然后将其作为特征，并基于其建立基于特征的索引结构来进行子图查询。属于这类的方法有 *GraphGrep,gIndex,FG-Index,Tree+Delta* 和 *GDIndex*。

在图索引中，最简单的特征就是通路 (walk)。一些情况下，我们可以用路径作为特征，就像这个领域的先驱方法 *GraphGrep*. 路径很好检索特性，并且很好处理。但是，路径的简单性却制约着这种方法的性能。举例而言，即便两图的所有路径都是不同长度的，我们利用路径也不能区分出环和链的拓扑结构。

当意识到路径的局限性后，*gIndex* 和 *FG-Index* 便应运而生了。这两种方法利用子图结构来有效分辨路径和环。但是，基于子图结构的索引也有一些限制，就是子图的枚举和匹配都是运算量非常大的过程。为了克服这些障碍，这些方法采用只提取频繁子结构来作为索引特征。还有些相似的方法，像 *Tree+Delta* 和 *TreePi* 则用频繁树结构来代替频繁子图结构来克服这些障碍。

*GDIndex* 也采用子图结构算法作为基本索引特征，但是并不使其局限于频繁子图结构，除此之外，这个算法还采用了哈希表来加速图同构查询。尽管这个算法设计目的是为了子图查询，但也同样支持相似查询。

## 1.2 模糊子图搜索

除了精确图匹配，也有些其他算法放宽了同构的限时，允许部分匹配或者模糊匹配。这是一个相对较新的领域，因此并没有太多的算法针对这个问题。有一个针对这个问题的算法，*SAGA*，它被设计用来作为生物路径分析。首先，建立一个基于图片段的索引。当候选图与查询图失配时，我们用一个图距离度量方法来度量差异。

另一个算法叫做 *gApprox* 和 *gIndex* 相似。从思想到名字，包括作者都很相似。这种算法力求从数据库中挖掘频繁模糊模式，并以此作为索引特征。他们还提出一个概念，称为模糊频繁。

TALE 算法也是用来做模糊匹配的一种算法。但是它着重于处理有上千节点的大图。

## 1.3 图相似性搜索

通常，我们有很多方法去测量图之间的相似度。第一种方法是*编辑距离 (edit distance)*. *编辑距离*就是我们将图 $G$ 通过一系列操作 (增删点边，重新标号等) 变换为另一个图 $G'$ 所需的操作数。我们可以通过给不同操作分配不同的费用，然后用费用总和当做距离来调整这个方法的准确度。虽然编辑距离是一种非常直观的图相似性测度方法，但是实际上我们难以计算它 (是个 NP-hard 问题).*C-Tree*[**?**] 是一种被广泛使用的图索引模型。它没有使用图的片段信息作为特征值，而是把数据图组织在一种内部节点是*图闭包 (graph closures)*, 叶节点是数据图的树形结构中。相比于前两种方法 *GraphGrep* 和 *gIndex*,*C-Tree* 的最大优势在于其支持相似性搜索，而前两个并不支持。

还有一种名为 *GString* 的子图相似性查询方法也和 *GraphGrep* 一样是用图片段作为特征值的。当然，这种方法与前面两种基于特征值的子图查询还是不同的。在这个方法中，首先我们分解复杂图为节点数较少的连通图，得到的这些连通图每个都是一个特定的图片段。随后，我们用一种标准的编号方式把数据库中的所有图都转化为一个个字符串。并用这些字符串构建一颗后缀树来支持相似性搜索。这种方法融合了子图的数据表达能力 (信息完整) 和用字符串匹配来查询图的速度 (速度快)。

另外, *最大公共子图 (maximal common subgraph)*[**?**] 和*图配对 (graph alignment)*[**?**, **?**] 这两种方法也常被用来定义图相似度。虽然有这么多方法，但是不幸的是迄今为止我们仍没有一种简单的方法来索引或者度量大图数据库。

## 1.4 图核函数 Graph Kernel Functions

目前业界有很多图核函数，而开创性的一个图核函数是 Haussler 在他对 *R 卷积核 (R-convolution kernel)* 的研究中提出的。现在大多数图核函数都遵循它提出的这种框架。R 卷积核是基于把离散的结构 (如图) 分解成一系列的组成元素 (子图) 这个思想的。我们可以定义许多这样的分解, 就像组成结构中的核心一样。R 卷积核保证无论选择怎样的分解方式或者结构核心，都能得要一个对称半正定的函数，或者一个结构间的核函数。这个关键性质将寻找离散结构核函数的问题简化为寻找分解方式和结构间的核函数。R 卷积核可以通过加权分解核来允许组件结构间的加权核。

目前图核函数的研究可以大致分为两类：第一类是考虑图中的所有可能组成结构（例如所有路径），然后以此来计算两图之间的相似度。这一类的算法有 product graph kernel,random walk based kernel 和基于点对间最短路径的核。第二类核函数是尝试通过一组特殊的 (有限的) 结构来计算局部相似度，并

且值在这有限的结构中统计共享的结构。这种方法包含一大类图核算法，叫做 spectrum 核，还有最近频繁子图核。我们发现最有效的核函数是 Vishwanathan 提出的用于全局相似性度量的方法。全局度量需要 $O(n^3)$，$n$ 是图中最大的节点数。众所周知，不同于全局相似度测量，局部相似性度量时间代价是非常昂贵的。因为子结构匹配（如子图同构）是一个 NP-hard 的问题。

我们采用一个最近提出的图小波匹配核，并将其扩展使其能在大数据库运行。