# 图数据中频繁模式挖掘算法研究综述

高 琳[1], 覃桂敏[1,2], 周晓峰[1]
(1. 西安电子科技大学计算机学院, 陕西西安 710071; 2. 西安电子科技大学软件学院, 陕西西安 710071)

摘 要: 本文对图数据中的频繁模式挖掘算法进行了综述. 依据算法的特性和数学基础对算法进行了分类, 主要集中于算法的求解思想和不同算法之间的关系的比较, 并对一些著名的算法进行了详细的分析和讨论. 基于算法的特性, 比较了各种算法适用的范围以及应用领域. 最后, 讨论了频繁模式挖掘的最新进展及未来的研究方向.

关键词: 频繁子图; 频繁模式挖掘; 图的匹配; 图的同构
中图分类号: TP391 文献标识码: A 文章编号: 0372-2112 (2008) 08-1603-07

# An Overview of Algorithms for Mining Frequent Patterns in Graph Data

GAO Lin[1], QIN Gui-min[1,2], ZHOU Xiao-feng[1]
(1. *School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi 710071, China;*
2. *Software School, Xidian University, Xi'an, Shaanxi 710071, China*)

**Abstract:** Graphs are widely used to model data in various applications. In recent years, a lot of algorithms have been developed to mine frequent pattern in graph data. We review some representative algorithms in this area. We classify the algorithms based on their properties and mathematical foundation, focusing on the algorithms underlying the approaches and how the algorithms relate to each other. Some typical algorithms are investigated. The comparison is then given based on the properties of the algorithms. New progress and future research directions are pointed.

**Key words:** frequent subgraph; frequent pattern mining; graph match; graph isomorphism

## 1 引言

在数据挖掘领域, 图数据中频繁模式的挖掘已经引起了人们极大的关注. 图挖掘已经成为数据挖掘和机器学习的新的研究方向. 而且, 图挖掘在许多实际应用中具有巨大的潜在价值, 这些应用包括语义网络[1]、行为建模[2]、生物网络分析[3~5]、化学化合物分类[6]和大分子分析[7]. 近年来, 基于图数据进行频繁子图挖掘算法的研究已经越来越引起人们的兴趣. 目前存在两类不同的图数据中频繁子图挖掘问题, 即图集(graph-transaction setting)和单图(single-graph setting)[8]. 这两类不同的问题导致了计算模式频率的方式也不同. 对于图集, 模式的频率由该模式在输入图集中出现的次数确定, 若在某个图中该模式出现多于一次, 则仅考虑一次; 对于单图, 模式的频率由该模式在一个图中出现的次数确定. 由于输入数据集以及问题定义的本质区别, 解决图集的算法不能用于解决单图问题, 而解决单图问题的算法可以很容易用来解决图集问题[8].

目前, 人们已经提出了很多高效的算法来挖掘图集中的模式[9~15], 也有一些算法用于发现单图中的模式[16~19], 一些算法用于约束图的挖掘[12,20,21,49,50]. 由于应用的对象及问题背景的不同, 这些算法的效果也具有极大的差异. 对于一个实际的问题, 面临的挑战是如何选择最有效的解决给定问题的算法. 因此, 有必要对这些算法进行分类. 本文首先综述了近几年来比较有代表性的一些算法, 依据算法的特性和数学基础对算法进行了分类, 并对一些著名的算法进行了详细的分析和讨论. 然后, 基于算法的特性, 比较了各种算法适用的范围以及应用领域. 最后, 讨论了频繁模式挖掘的最新进展及未来的研究方向.

## 2 定义和理论基础

本节回顾图论相关的概念, 给出图的规范化标记方法, 并讨论图集和单图中子图频率的定义方式.

### 2.1 图的相关定义[22]

对图 $G=(V,E)$, 如果一个图中每两个顶点之间都有路径, 则称该图是连通的.

给定图 $G=(V,E)$, $G_S=(V_S,E_S)$, 当且仅当 $V_S\subseteq V$

收稿日期: 2007-07-30; 修回日期: 2007-09-26
基金项目: 国家自然科学基金(No. 60574039); 教育部留学回国人员启动基金; 西安应用材料基金(No. XA-AM-200605)

且 $E_S \subseteq E$, 称图 $G_S$ 为图 $G$ 的子图. 同样, 当且仅当 $V_S \subseteq V$ 且 $E_S$ 包含 $E$ 中的所有连接 $V_S$ 中的顶点的边, 称 $G_S$ 是 $G$ 的导出子图.

如果这两个图 $G_1=(V_1, E_1)$ 和 $G_2=(V_2, E_2)$ 在拓扑结构上是相同的, 则称两个图是同构的, 即存在一个 $V_1$ 到 $V_2$ 的对应关系, 使得 $E_1$ 中的每一条边对应 $E_2$ 中唯一的一条边, 反之亦然. 子图同构问题是指是否存在 $G_2$ 和 $G_1$ 的某个子图同构, 也就是说判断是否 $G_2$ 为 $G_1$ 所包含.

**2.2** 图的规范化标记

图 $G$ 的规范化标记 $d(G)$ 是指图 $G$ 的一个唯一编码(如字符串), 该编码由图 $G$ 的拓扑结构唯一确定与图 $G$ 的顶点和边的顺序无关. 因此, 如果两个图同构, 则这两个图具有相同的规范化标记. 规范化标记以唯一确定的方式建立一组图的完整顺序, 并且可以用于两个图的快速比较. 规范化标记在频繁子图发现算法中起着非常重要的作用. 然而, 确定一个图的规范化标记的问题复杂度很高, 等价于确定两个图之间的同构问题, 因为如果两个图同构, 则它们的规范化标记必须相同[14].

**2.3** 子图频率的计算

输入图的集合 $D=\{G_1, G_2, \cdots, G_n\}$, 和一个参数 $\sigma$ $(0<\sigma\leq 1)$. 频繁子图发现的目的是找出所有连通的子图, 这些子图至少出现 $\sigma|D|$ 次, 其中 $|D|$ 是图集 $D$ 的势, $\sigma$ 是支持度阈值.

对于单图, 输入是一个图 $G$, 和其它一些条件如频率阈值 $f$(找出至少在图 $G$ 中出现 $f$ 次的子图). 根据图的哪些元素(顶点或边)被不同的嵌入共享, 存在不同的子图频率的定义方式. 表 1 列出了四种可能的定义方式, 其中只有三种是可行的[23]. 频繁子图挖掘的目标是找出所有满足上下文给定条件的连通子图.

表 1 频率定义方式

| 定义方式 | 不同嵌入可以共享的图的元素 | | 不同的情况 |
|---|---|---|---|
| | 顶点 | 边 | |
| $F1$ | 是 | 是 | 可行 |
| $F2$ | 是 | 否 | 可行 |
| $F*$ | 否 | 是 | 不可行 |
| $F3$ | 否 | 否 | 可行 |

## 3 频繁模式挖掘算法

基于图的频繁子图挖掘算法按照应用对象可以分为两类: 图集和单图. 本节依据算法的特性及相关的数学基础分类介绍这些算法, 并分析其中最有代表性的算法.

**3.1** 图集中的频繁子图挖掘算法

这些算法包括基于贪心搜索的方法、基于归纳逻辑程序设计(ILP)的方法和基于图论的方法.

**3.1.1** 基于贪心搜索的方法

基于贪心搜索策略进行频繁子图的挖掘, 在 1994 年取得了两个开创性的成果 SUBDUE 和 GBI. 其中最著名的是 SUBDUE 算法[24]. SUBDUE 基于最小描述长度原则(MDL), 通过用一个顶点替换模式来找出那些可以有效压缩原始输入数据的模式. 算法从仅包含输入图 $G$ 中的一个顶点的子图开始, 通过每次增加一个顶点来扩展子图. SUBDUE 的一个优势是它可以进行子图的近似匹配. 而且它还可以以预定义子图的形式来嵌入背景知识. SUBDUE 采用一种启发式的束(beam)搜索方法来缩小搜索空间, 以提高计算性能. 后来 SUBDUE 还扩展成为图分类算法, 称为 SubdueCL[26]. 在 SubdueCL 中不再采用最小描述长度, 而是采用基于子图置信度的启发式方法. 另一种基于启发式方法的算法是 MOLFEA[27], 它充分利用了化合物的 SMILES 字符串表示, 识别与频繁出现的子序列相对应的子结构.

算法 GBI[17] 类似于 SUBDUE. GBI 用一个顶点代替每个发现的子图来不断的压缩图, 以得到具有最小规模的图. 它采用经验图规模的定义, 反映了压缩图和提取模式的规模. 这样搜索的优点是妨碍了持续压缩. GBI 可以处理带有闭路径(包括环)的有向或者无向标签图. 搜索过程中的每一步基本操作是通过一条边或者块找到好的连接的顶点对集合, 采用规范化标记来判断得到的子图是否同构[28]. GBI 还作为特征构造器, 用于构造图数据中的决策树分类器的特征[29].

**3.1.2** 基于 **ILP** 的方法

图可以很容易用一阶逻辑[14]来表示, 于是提出了基于 ILP 系统的挖掘频繁子图的方法[30~34]. 基于 ILP 算法的目标是归纳出一个可以正确的分类正样本集和负样本集的规则集. 在 ILP 系统中建立图模型的情况下, 这些规则通常对应于子图. 大多数基于 ILP 的方法本质上是贪心的, 采用不同的启发式方法来剪枝可能的假设空间. 因此, 它们趋向于找出具有高支持度的子图, 并且可以作为较好的类识别器. 然而, 它们不能保证可以发现所有的频繁子图. 有一个例外是 Dehaspe 等提出的 ILP 系统 WARMR[35], 该系统可以发现所有的频繁子图, 然而, WARMR 并不是专门为了处理图结构而设计的, 它也不采用图模型特定的优化技术, 因此它具有很高的计算量. 另一个例子是 WARMR 系统[36,37], 它是专门为发现所有可能的出现次数不小于某个特定的频率阈值的化学化合物子结构而提出的. 然而, WARMR 的计算复杂度很高, 因此只能用于发现出现次数相对较高的子结构.

**3.1.3** 基于图论的方法

AGM 是最早的基于图论的方法. 最初, AGM(基于 Apriori 的图挖掘)只能发现频繁导出子图[38], 后来该算

法得到扩展可以发现所有的频繁子图[39]. 算法采用广度优先搜索(BFS)方法, 每次迭代增加一个顶点来扩展频繁子图. 为了区分不同的子图, 算法采用基于邻接矩阵表示法的规范化标记来表示图. AGM 算法可以获得较好的性能[38]. 算法的改进[39]通过保存原先已找到的频繁模式的嵌入来减少子图同构计算, 以更大的内存需求来获得算法性能的极大提高. AGM 算法可以挖掘各种不同的子图, 包括一般的子图、导出子图、连通子图、有序子树、无序子树和子路径.

Borgelt 等[40]提出的挖掘化学子结构的算法 MoFa 采用深度优先搜索(DFS)方法, 类似于 Eclat[41]提出的发现频繁项集的方法. 在该算法中, 一旦识别了一个频繁子图, 就在包含该子图的输入数据集中扩展该频繁子图. 为了减少子图同构的计算, 该算法保存了前面发现的子图的嵌入, 并通过类似于 AGM[39]的策略增加一条边来扩展嵌入. 而且, 由于频繁子图的所有嵌入均是已知的, 通过删除不在任意嵌入中的边和顶点来把原数据集映射到一个更小的集合. 尽管采用了这些优化策略, 该算法的效率仍比 FSG(频繁子图发现)[42, 14]低. 有两个原因. 第一, 这些算法的候选子图生成方法不能保证一个子图只产生一次; 第二, 在化学数据集中同样的子图有许多个嵌入. FSG 的主要特征是: 每次通过增加一条边来扩展频繁子图的规模; 在候选子图生成和频率计数中采用了多种优化策略使得该算法可以用于大规模的数据集; 并且它采用复杂的算法计算规范化标记以唯一识别各种生成子图, 而不采用复杂的图和子图同构计算方法[14]. gSpan[9]采用深度优先搜索方法发现频繁子图. 与 MoFa[40]算法不同, gSpan 每次生成一个候选子图时, 就计算其规范化标记. 如果该规范化标记是最小的, 则保留下来进行进一步的深度搜索的扩展. 如果不是最小的, 则丢弃该候选子图, 因为一定存在到达该候选子图的另一条路径. 因此, gSpan 避免了冗余候选子图的产生. 为了保证能有效的比较这些子图, 该算法采用了深度优先遍历的规范化标记方法. 而且, gSpan 并不保存所有原先发现的频繁子图的嵌入的信息, 因此大大节省了内存的使用. 然而, 其实时识别所有的嵌入并将这些嵌入映射到数据集的方法与文献[38]中采用的方法是相似的. 根据文献[9]中的性能报告, gSpan 和 FSG 在 PTE 数据集上不相上下, 而在合成数据集上, gSpan 比 FSG 的性能要好. Huan 等提出的 FFSM(快速频繁子图挖掘)[43]通过有效的处理子图同构问题并结合一个可以减少候选子图生成的代数图框架, 提高了算法性能. 通过不同的数据集进行极限测试, 结果表明 FFSM 比 gSpan 的性能更优. 与大部分现有的算法相比, PATH[10, 44, 15]是一种基于 Apriori 的频繁子图发现算法, 它可以发现有向图模式、无向图模式、包含圈的模式. 增加的功能可以挖掘在线文档中通过引用连接起来的块, 并且可以支持无向图. 在搜索频繁模式过程中, PATH 采用频繁路径构造候选子图. 经验表明该算法比其它算法有明显的优势. 该算法可以以几种方式扩展, 如采用部分标记模式, 采用更复杂的块(树), 调整算法使其适于动态数据库模型, 以及采用 Apriori-TID 技术[15]. Nijssen 和 Kok[45, 46]提出了算法 Gaston, 该算法使用图编码和子图同构, 枚举出分子集中所有的频繁导出亚分子.

gSpan 算法是一个典型的频繁子图挖掘算法, 该算法的核心思想和算法步骤如下[9]:

(1) 用 DFS 编码来表示一个图: 由于一个图可能对应多个 DFS(深度优先搜索)树, 对每个 DFS 树, 根据顶点的访问顺序和某些规则转换为边序列, 称为 DFS 编码.

(2) 选择最小的 DFS 编码来唯一表示一个图: 每个图对应唯一的最小 DFS 编码, 因此图的同构计算可以通过比较其最小 DFS 编码来进行.

(3) 产生候选子图: 按最右扩展规则来扩展子图, 对 k 阶频繁子图的 DFS 编码树进行最右路径扩展, 每次添加一条边, 得到 $k+1$ 候选子图. 图 1 是上述 gSpan 搜索过程形成的搜索树, 树中每个节点是 DFS 编码, 第 $n$ 层中节点为第 $n-1$ 层节点扩展而得. 图中若 $s$ 和 $s'$ 表示同一个图, 则可以将 $s'$ 剪枝.

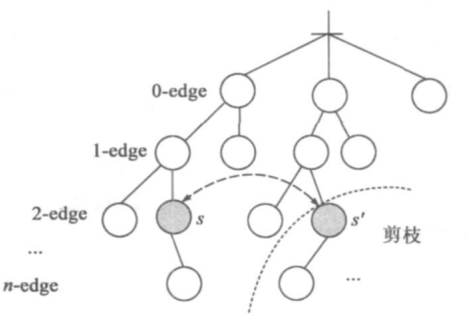


图 1　搜索树

(4) 剪枝: 如果 $k+1$ 候选子图不是最小编码形式的, 则认为该图是冗余的, 从候选子图中删除.

(5) 每次在计算 $k$ 阶频繁子图的支持度的时候, 同时记录频繁子图的所有嵌入. 这样, $k+1$ 阶候选子图的支持度就可以通过对 $k$ 频繁子图的所有嵌入进行最右路径扩展获得.

算法步骤如下:

输入: DFS 编码 $s$, 图集 $D$, 最小支持度 min*Sup*

输出: 频繁子图集 $S$

Subgraph_Mining($D$, $S$, $s$);

Step1: if $s \neq \min(s)$　resturn;

Step2: $S \leftarrow S \cup \{s\}$;

Step3: enumerate $s$ in each graph in $D$ and count its

children;

Step4: for each $c$, $c$ is $s'$ child do

Step5: if $support(c) \geq \min Sup$

Step6: $s \leftarrow c$;

Step7: Subgraph_Mining($D_s$, $S$, $s$);

### 3.2 单图中的频繁子图挖掘算法

SUBDUE 和 GBI 不仅用于图集的频繁子图挖掘,而且是最著名的单图频繁子图挖掘算法. 2002 年 Ghazizadeh 等[18]最早针对单图提出了 SEuS 算法,主要思想是先根据顶点编号对输入图进行有损压缩得到一种叫做 summary 的数据结构, summary 有助于快速的修剪掉不频繁的候选子图. 当图中含有很少的子图种类却有很高的频率时,这种压缩源数据的方法有很好的效果,在相反情况下效率却很低. 值得注意的是由于是有损压缩, SEuS 算法和前面的两个算法一样是不精确的. Vanetik 等[47]提出了一种基于边的子图增长策略的算法,适用于有标记的无向单图,并且可以精确挖掘出所有的频繁子图. 该算法中各个嵌入间边不可重叠,每类子图的嵌入个数就是该子图的发生频率. 2005 年 Kuramochi 等[8]提出了 SiGram(Pafi) 算法,作者提出了子图间边重叠的频繁子图挖掘问题,但并没有提出相应的解决算法,该算法利用边不可重叠子图的反向闭包性质,采用广度和深度两种增长策略在子图的精确挖掘中达到了很好的效果. 可重叠子图的挖掘在后基因组分子生物学等领域有非常重要的意义,有助于从海量信息的生物网络数据中提取出基因间的相互关系和功能模块. 我们提出了一种新的算法可以对单图数据进行允许边重叠的频繁子图挖掘,通过对多个生物网络数据的仿真实验证明该算法达到了很好的效果,请参阅文献[49].

### 3.3 基于约束的图模式挖掘

前面讨论的频繁子图挖掘处理带标记或不带标记、无向的或有向的、没有任何约束的简单连通图. 然而,许多应用需要对要挖掘的模式施加各种约束,或者寻找不同的子结构模式. 例如,可能希望挖掘的每个模式包含某些特殊的顶点或边,或者顶点或边的总数在指定的范围内,或者寻找图模式的平均密度超过一个阈值的模式[45]. 本节,我们研究基于约束的图模式挖掘.

#### 3.3.1 特殊的子图挖掘

频繁子图的挖掘时间复杂度非常高,为了尽量减少运算时间,需要根据不同的应用需求,对子图加以限制.

(1)闭频繁子图

一个频繁图 $G$ 是闭图当且仅当不存在于 $G$ 有相同支持度的真超图 $G'$. 在相同的最小支持度阈值下,闭子图模式集和子图模式全集具有相同的表达能力,因为后者能够被闭包模式的推导集产生. 与普通的频繁子图挖掘相比,挖掘闭图将有相似的精确性,有低冗余和高效率. Yan 和 Han 在文献[12]中提出了一种称作 CloseGraph 的高效方法,它通过 gSpan 算法的扩展挖掘闭频繁图. 实验研究已经表明 CloseGraph 经常产生更少的图模式,并且比挖掘全部模式集的 gSpan 更有效.

(2)凝聚频繁子图

频繁子结构 $G$ 是一个凝聚子图(coherent subgraph),如果 $G$ 和它的每个子图之间的互信息都超过某个阈值. 凝聚子结构的数量显著少于频繁子结构的数量,挖掘凝聚子结构能够有效剪除冗余模式.

(3)稠密频繁子图

存在一种称作关系图的特殊图结构,这种图的每个节点在每个图中仅用一次. Hu 等[20]提出了一种高效挖掘凝聚稠密频繁子图的算法. Yan 等[21]提出了两种算法挖掘闭稠密频繁子图.

(4)哈密尔顿环频繁子图

在对生物网络功能模体发现和化学制药的药物发现中,一个关键步骤就是在给定的网络中挖掘出特定结构的频繁子图. 注意到在已经发现的模体中都有哈密尔顿环的性质. 因此我们基于矩阵论的方法挖掘哈密尔顿环频繁子图,请参阅文献[50].

#### 3.3.2 基于约束的子结构模式挖掘

在频繁子图挖掘问题中,不同的应用领域需要给出不同的约束条件[45].

(1)元素、集合或子图包含约束. 这种约束可以取给定的子图集作为查询,首先使用该约束进行选择,然后对选取的数据集从给定的子图及增长模式进行挖掘.

(2)几何约束. 几何约束是每对互相连接的边之间

表 2 频繁子图挖掘算法比较

| 算法 | 子图类型 | 遍历策略 | 数据集 | 发表时间(年) |
|---|---|---|---|---|
| SUBDUE | 连通子图 | 贪心策略 | 图集/单图 | 1994- 2000 |
| GBI | 连通子图 | 贪心策略 | 图集/单图 | 1994- 2000 |
| WARMR | 连通子图 | 归纳逻辑程序设计 | 图集 | 1990- 2001 |
| AGM | 导出子图 | 广度优先 | 图集 | 2000- 2003 |
| AcGM | 连通子图 | 广度优先 | 图集 | 2002 |
| MoFa | 连通子图 | 深度优先 | 图集 | 2002 |
| SEuS | 连通子图 | 贪心策略 | 单图 | 2002 |
| gSpan | 连通子图,树,路径 | 深度优先 | 图集 | 2002- 2003 |
| FFSM | 连通子图 | 广度优先 | 图集 | 2003 |
| CloseGraph | 闭子图 | 深度优先 | 图集 | 2003 |
| FSG | 连通子图 | 广度优先 | 图集 | 2001- 2004 |
| GASTON | 导出子图 | 深度优先 | 图集 | 2004 |
| SiGraM(pafi) | 不可重叠连通子图 | 广度优先和深度优先 | 单图 | 2004 |
| PATH | 连通子图 | 广度优先 | 图集/单图 | 2002- 2006 |
| gPrune[51] | 连通子图 | 广度优先和深度优先 | 图集 | 2007 |
| GraphGen[54] | 连通子图 | 深度优先 | 图集 | 2007 |

的角必须在一定的范围内的约束.

(3) 值和约束. 例如, 这种约束可以是边上的(正的)权重之和 $Sum_e$ 在 low 和 high 之间.

还有一些其它情况, 比如在不完全标记图中的频繁子图挖掘, 或者输入图不是简单图(点有自环, 或者一对点之间不止一条边), 或者在非连通图中挖掘非连通的频繁子图, 或者在树形结构中挖掘频繁出现的子树.

### 3.4 图挖掘算法的比较与分析

频繁子图挖掘算法有以下一些重要性质[50]: (1) 搜索顺序, 深度优先或者广度优先; (2) 候选子图的产生策略, Apriori 或者 Pattern-growth; (3) 对复制图的消除策略, 主动地或被动地; (4) 支持度的计算方法; (5) 模式的发现顺序. 大部分的 BFS 算法都采用了 Apriori 的候选子图产生策略. 在表 2 中, 详细列出了一些频繁子图挖掘算法的重要性质, 并进行了比较.

## 4 结束语

随着生物信息学, 化学情报学, 计算机视觉, 视频索引, Web 分析等领域的飞速发展, 图数据作为一种数据结构在复杂问题的建模中变得越来越重要, 为了进一步对图进行特征化、区分、分类和聚类分析, 频繁子图挖掘已经成为一项重要的任务. 本文综述了典型的频繁子图挖掘算法, 并主要介绍了这些算法的应用领域及这些算法相互间的联系. 此外, 分析了这些算法的优势及不足之处. 从理论角度看, 频繁子图挖掘算法在图特征和同构等方面还有很多问题. 这个领域在应用和理论上都有很高的研究价值, 已经成为数据挖掘的一个重要研究方向. 从 1994 年至今, 这个领域已经有数百篇论文发表, 大量的科学研究和应用开发层出不穷. 由于篇幅有限, 本文不可能涵盖这个领域的所有内容. 希望这篇综述能对频繁子图挖掘算法的研究起到一定的参考作用.

参考文献:


[1] B Berendt, A Hotho, G Stumme. Towards semantic web mining [A]. Ian Horrocks and James Hendler. In International Semantic Web Conference (ISWC) [C]. Sardinia, Italy: Springer-Verlag, 2002. 264-278.

[2] M Girvan, M E J Newman. Community structure in social and biological networks[J]. Proc Natl Acad Sci, 2001, 99(12): 7821-7826.

[3] N Kashtan, S Itzkovitz, R Milo, U Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs[J]. Bioinformatics, 2004, 20(11): 1746-1758.

[4] N Przulj, D G Corneil, I Jurisica. Efficient estimation of graphlet frequency distributions in protein-protein interaction networks [J]. Bioinformatics, 2006, 22(8): 974-980.

[5] M Koyuturk, Y Kim, S Subramaniam, et al. Detecting conserved interaction patterns in biological networks[J]. Journal of Computational Biology, 2006, 13(7): 1299-1322.

[6] M Deshpande, M Kuramochi, G Karypis. Frequent substructure based approaches for classifying chemical compounds[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(8): 1036-1050.

[7] S Parthasarathy, M Coatney. Efficient discovery of common substructures in macromolecules[A]. Proc of 2002 IEEE International Conference on Data Mining (ICDM) [C]. Maebashi City, Japan: IEEE, 2002. 362-369.

[8] M Kuramochi, G Karypis. Finding frequent patterns in a large sparse graph[J]. Data Mining and Knowledge Discovery, 2005, 11(3): 243-271.

[9] X Yan, J Han. gSpan: Graph-based substructure pattern mining [A]. Proc of 2002 IEEE Int'l Conf. Data Mining (ICDM) [C]. Maebashi City, Japan: IEEE, 2002. 721-724.

[10] A Inokuchi, T Washio, H Motoda. Complete mining of frequent patterns from graphs: Mining graph data[J]. Machine Learning, 2003, 50(3): 321-354.

[11] M Hong, H Zhou, W Wang, B Shi. An efficient algorithm of frequent connected subgraph extraction[A]. Proc of the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-03) [C]. Volume 2637 of Lecture Notes in Computer Science. Seoul, South Korea: Springer-Verlag, 2003. 40-51.

[12] X Yan, J Han. CloseGraph: Mining closed frequent graph patterns[A]. Proc of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003) [C]. Washington, DC, USA: ACM, 2003. 286-295.

[13] J Huan, W Wang, J Prins. Efficient mining of frequent subgraph in the presence of isomorphism[A]. In Proc of 2003 IEEE International Conference on Data Mining (ICDM'03) [C]. Melbourne, Florida USA: IEEE, 2003. 549-552.

[14] M Kuramochi, G Karypis. An efficient algorithm for discovering frequent subgraphs[J]. IEEE Trans. Knowledge and Data Eng, 2004, 16(9): 1038-1051.

[15] E Gudes, S E Shimony, N Vanetik. Discovering frequent graph patterns using disjoint paths[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(11): 1441-1456.

[16] L B Holder, D J Cook, S Djoko. Substructure discovery in the SUBDUE System[A] Proc AAAI Workshop Knowledge Discovery in Databases[C]. Seattle, WA: AAAI Press, 1994. 169-180.

[17] K Yoshida, H Motoda, N Indurkhya. Graph-based induction as a unified learning framework[J]. Journal of Applied Intelligence, 1994, 4: 297-328.

[18] S Ghazizadeh, S Chawathe. SeuS: Structure extraction using summaries[A]. Proc of the 5th Intl. Conf. on Discovery Sci-

ence[C]. Lübeck, Germany: ACM, 2002. 71- 85.

[19] N Vanetik, E Gudes, S E Shimony. Computing frequent graph patterns from semistructured data[A]. In Proc of 2002 IEEE International Conference on Data Mining (ICDM)[C]. Maebashi City, Japan: IEEE, 2002. 458- 465.

[20] Haiyan Hu, Xifeng Yan, Yu Huang, Jiawei Han, Xianghong Jasmin Zhou. Mining coherent dense subgraphs across massive biological networks for functional discovery[J]. Bioinformatics, 2005, 21(1): i213- i221.

[21] X Yan, X J Zhou, J Han. Mining closed relational graphs with connectivity constraints[A]. Proc. 2005 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD' 05)[C]. Chicago, IL, USA: ACM, Aug. 2005. 324- 333.

[22] W T Tutte. Graph Theory[M]. Cambridge University Press, 2001.

[23] F Schreiber, H Schwobbermeyer. Towards motif detection in networks: Frequency concepts and flexible search[A]. NETTAB' 04[C]. Camerino, Italy, 2004. 91- 102.

[24] J Cook, L Holder. Substructure discovery using minimum description length and background knowledge[J]. J Artificial Intelligence Research, 1994, 231- 255.

[25] D J Cook, L B Holder. Graphr based data mining[J]. IEEE Intelligent Systems, 2000, 15(2): 32- 41.

[26] J Gonzalez, L Holder, D Cook. Application of graph based concept learning to the predictive toxicology domain[A]. PTC, Workshop at the 5th PKDD[C]. Freiburg, Germany: Springer Verlag, 2001.

[27] S Kramer, L De Raedt, C Helma. Molecular feature mining in Hiv data[A]. 7th International Conference on Knowledge Discovery and Data Mining[C]. San Francisco, California, USA: ACM, 2001. 136- 143.

[28] T Washio, H Motoda. State of the art of graphr based data mirring[J]. ACM SIGKDD Explorations Newsletter 5, 2003: 59- 68.

[29] W Geamsakul, T Matsuda, T Yoshida, H Motoda, T Washio. Classfier construction by graphr based induction for graphr structured data[A]. Proc. of 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD' 03)[C]. Seoul, KOREA: Springer Verlag (LNAI 2637), 2003. 52- 62.

[30] J R Quinlan. Learning logical definitions from relations[J]. Machine Learning, 1990, 5: 239- 266.

[31] S Muggleton, L DeRaedt. Inductive logic programming: Theory and methods[J]. J Logic Programming, 1994, 19(2): 629- 679.

[32] S H Muggleton. Inverse entailment and progol[J]. New Generation Computing, special issue on inductive logic programming, 1995, 13(3): 245- 286.

[33] S H Muggleton. Scientific knowledge discovery using inductive logic programming[J]. Comm. ACM, 1999, 42(11): 42- 46.

[34] L Dehaspe, L De Raedt. Mining association rules in multiple relations[A]. Proc. Seventh Int' l Workshop Inductive Logic Programming[C]. Prague, Czech Republic: Springer Verlag, 1997. 125- 132.

[35] D K Ross, A Srinivasan, L Dehaspe. WARMR: A data mining tool for chemical data[J]. Journal of Computer Aided Molecular Design, 2001, 15: 173- 181.

[36] L Dehaspe, H Toivonen, R D King. Finding frequent substructures in chemical compounds[A]. In R. Agrawal, P. Stolorz, and G. Piatetsky-Shapiro. 4th International Conference on Knowledge Discovery and Data Mining[C]. New York, USA: AAAI Press, 1998. 30- 36.

[37] A Inokuchi, T Washio, H Motoda. An apriorr based algorithm for mining frequent substructures from graph data[A]. Proc. Fourth European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD' 00)[C]. Lyon, France: Springer Verlag, 2000. 13- 23.

[38] A Inokuchi, T Washio, K Nishimura, H Motoda. A fast algorithm for mining frequent connected subgraphs, Technical Report RT0448[R]. IBM Research, Tokyo Research Laboratory, 2002.

[39] C Borgelt, M R Berthold. Mining molecular fragments: finding relevant substructures of molecules[A]. Proc. 2002 IEEE Int' l Conf. Data Mining (ICDM)[C]. Maebashi City, Japan: Springer Verlag, 2002. 51- 58.

[40] M J Zaki, K Gouda. Fast vertical mining using diffsets, Technical Report 01-1[R]. Dept. of Computer Science, Rensselaer Polytechnic Inst, 2001.

[41] M Kuramochi, G Karypis. Frequent subgraph discovery[A]. Proc. 2001 IEEE Int' l Conf. Data Mining (ICDM)[C]. San Jose, California, USA: Springer Verlag, 2001. 313- 320.

[42] N Vanetik, E Gudes. Mining frequent labeled and partially labeled graph patterns[A]. Proc. Int' l Conf. Data Eng. (ICDE' 04)[C]. Boston, USA: Springer Verlag, 2004. 91- 102.

[43] S Nijssen, J N Kok. A quick start in frequent structure mining can make a difference[A]. Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004)[C]. Seattle, WA, USA: Springer Verlag, 2004. 647- 652.

[44] S Nijssen, J N Kok. Frequent graph mining and its application to molecular databases[A]. Proc. IEEE Int' l Conf. Systems, Man, and Cybernetics[C]. New York, USA: Springer Verlag, 2004. 4571- 4577.

[45] Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques[M]. 2nd ed. MORGAN KAUFMANN, 2006.

[46] Xiaofeng Zhou, Lin Gao, Anguo Dong. An algorithm for finding frequent patterns in a large sparse graph[A]. IAENG International Conference on Bioinformatics (ICB2007)[C]. Hong

Kong: IAENG, 2007. 290– 294.

[47] An guo Dong, Lin Gao, Xiaofeng Zhou. An algebra approach for finding frequent subgraphs with Hamilton cycle[A]. The 4th International Conference on Fuzzy Systems and Knowledge Discovery(FSKD'07)[C]. Haikou: IEEE, 2007. 288– 292.

[48] Feida Zhu, Xifeng Yan, Jiawei Han, Philip S. Yu. gPrune: A constraint pushing framework for graph pattern mining[A]. Proc. 2007 Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'07)[C]. Nanjing: Springer, 2007. 388 – 400.

[49] Jiawei Han, Hong Cheng, Dong Xin, Xifeng Yan. Frequent pattern mining: Current status and future directions[J]. Data Mining and Knowledge Discovery, 2007, 14(1): 55– 86.

[50] Xifeng Yan. Mining, Indexing, Similarity Search in Large Graph Data Sets[D]. University of Illinois at Urbana-Champaign, USA, 2006.

[51] 李先通, 李建中, 高宏. 一种高效频繁子图挖掘算法[J]. 软件学报, 2007, 18(10): 2469– 2680.
Li Xian-Tong, Li Jiang-Zhong, Gao Hong. An efficient frequent subgraph mining algorithm[J]. Journal of Software, 2007, 18(10): 2469– 2680. (in Chinese)

作者简介:

高　琳　女, 1964 年 11 月生, 西安电子科技大学计算机学院教授, 博士生导师, 学术带头人. 西安交通大学数学系计算数学专业理学学士学位、西北大学数学系计算数学专业理学硕士学位、西安电子科技大学电子工程研究所电路与系统专业博士学位. 2004 年 6 月至 2005 年 6 月被国家留学基金委批准选派赴加拿大 University of Guelph 做访问学者, 从事计算生物信息学交叉学科的研究工作. 近年来, 主要研究方向包括计算生物信息学、生物数据挖掘、图论与组合优化算法及其应用等, 在国内外核心期刊和国际会议发表学术论文 40 余篇. E-mail: lgao@mail.xidian.edu.cn

覃桂敏　女, 1977 年 1 月生, 西安电子科技大学软件学院讲师, 现为西安电子科技大学计算机学院在职博士. 西安电子科技大学机电工程学院学士、计算机学院硕士. 研究方向为计算生物信息学、算法设计与分析, 目前感兴趣的问题为频繁子图的挖掘、生物网络数据的模式发现算法研究.