基于图的频繁子结构挖掘算法综述*

张焕生[1]，崔炳德[1]，王政峰[1]，徐德生[2]
(1.河北工程技术高等专科学校 计算机应用技术 河北 沧州 061001;
2.北京科技大学 信息工程学院 北京 100083)

摘 要：随着对大量结构化数据分析需求的增长，从图集合中挖掘频繁子图模式已经成为数据挖掘领域的研究热点。通过对目前有代表性的频繁子图挖掘算法的分析和比较，全面总结了各算法的特性及优缺点，并预测了今后的发展趋势。

关键词：数据挖掘；频繁子结构；子图同构
中图分类号：TP391.75 文献标识码：A

Review of graph-based frequent substructure mining algorithm

ZHANG Huan Sheng[1], CUI Bing De[1], WANG Zheng Feng[1], XU De Sheng[2]
(1.Computer Application Technology, Hebei Engineering and Technical College, Cangzhou 061001, China;
2.School of Information Engineering, University of Science and Technology Beijing, Beijing 100083, China)

Abstract: With the increasing demand of massive structured data analysis, mining frequent subgraph patterns from graph datasets has been an attention-deserving field. This paper fully summarizes the characteristic and the advantages and disadvantages of these algorithms by analysis and comparison of popular frequent subgraph mining algorithm at present, and points out the future development trend.

Key words: data mining; frequent substructure; subgraph isomorphism;

随着生物信息学(蛋白质结构、基因组识别和比较分析)、社会网络(实体间的联系)、Web分析(Web的链接结构分析、Web内容挖掘和Web日志的搜索)以及文本信息检索(文档的选择、文档的秩评定)等复杂结构的广泛应用。图作为一般数据结构在这些结构以及它们的建模方面日趋重要。为了进一步对图进行特征化、区分、分类和聚类分析，挖掘频繁子图模式已经成为一项重要的任务，日益受到人们的关注。

1 现有的频繁子图挖掘方法

在各种各样的图模式中，频繁子结构是可以在图集合中发现的非常基本的模式。在大型图数据库中可以用它建立图索引并进行相似性搜索，区分不同的图组群，对图进行分类和聚类分析。目前已经有了一些成熟的频繁子结构的挖掘方法，并且在许多领域得到了应用，尤其在药物发现和化合物合成领域的应用更为流行，目前子结构挖掘算法分类如下：

(1)基于Apriori的频繁结构挖掘算法：AGM、PSG、路径连接算法；

(2)基于模式增长的频繁结构挖掘算法：Espan、FFSM、CloseGraph；

(3)基于最小描述长度的近似频繁子结构挖掘算法：SUBOUE；

(4)基于模式增长和模式归约的精确稠密频繁子结构挖掘算法：CloseCut、Splat。

1.1 基于Apriori的频繁子结构挖掘算法

基于Apriori的频繁子结构挖掘算法与基于Apriori的频繁项集挖掘算法相类似。频繁图的搜索开始于小规模图，按照自底向上的方式产生具有附加顶点、边或路

*基金项目：“十一五”国家科技支撑课题“基于认知的名老中医学术思想临床经验挖掘技术研究”

径的候选图。最近提出的基于Apriori的频繁子结构挖掘算法包括AGM、FSG和路径连接方法等。

（1）由Inokuchi等人提出的计算高效性算法AGM[1]，能找到所有满足某一最小支持度阈值的频繁子图，它与基于Apriori的项集挖掘算法具有类似的特点，使用基于顶点的候选产生方法，通过在每一步增加一个顶点来扩展子结构的规模。2个大小为k的频繁图进行连接，仅当它们具有相同的大小为k–1的子图。新形成的候选包括1个大小为k–1的公共子图和来自2个大小为k的模式中的2个附加顶点。AGM能够处理带有标记的顶点和边的图，可以有效地挖掘不同类型的子图，例如一般子图、导出子图、连通子图、有序子树、无序子树和子路经，特别对于合成密集型数据集具有良好的性能。这种方法常用于发生变异活动的化合物分子结构的分析。实验证明，在一个包含300个化合物的数据集中，当最小支持度阈值从20%～10%变化时要找出所有的频繁生成子图需要40 min到8天的时间。AGM对于一个生成子图可以是非连通的，包含几个独立的图片段，但这种方法也需要很长的处理时间。

（2）Kuramochi等人利用边增长策略进一步发展了上述思想，提出了FSG算法[2]。该算法在具有多条边和顶点标记的图数据集中能更好地运行，运行时间依赖于被发现的频繁子图的大小。它的输入为图集，但为了减小时间复杂度，FSG限用于连通图，由于很多应用都可以转化为连通图，所以FSG的这个限制并未影响到它的应用范围。为了提高导出规范标记效率，它使用了一些图顶点不变量(例如设定图中的每个顶点的度)并且它通过引入TID(transaction ID)方法提高了频繁子图的候选产生效率。此外它采用基于边的候选产生策略，通过每次增加一条边来扩展子结构的规模。合并2个大小为k的频繁子图，当且仅当他们共享相同的具有k–1条边的连通主子图(该主子图称为核)，新形成的侯选包括核和来自2个大小为k的模式中的2条附加边，通过这种合并方法还提高了FSG的连接效率。正因为FSG引入了这些技术所以运行速度很快，实验证明，它具有良好的性能并且能够随数据库的大小呈线性比例变化，它能够从一个包含8万个图片支持度阈值为2%的合成数据集中，以少于500 s的时间发现所有的频繁连通子图。但是对于大型图数据存储TID列表要占用大量的内存空间，而且不像合并项集那样，2个大小为k的频繁项集能够只产生唯一的大小为k+1的项集，这里2个大小为k的子图可能产生多个大小为k+1的候选子图，生成大量的重复候选子图，降低了算法的整体效率。

（3）边不相交路径方法[3]依据图所拥有的不相交路径的数量来分类。如果2条路径不共享任何边，那么就称这两条路径是边不相交的[4]。该方法提出一种合成关系(composition relation)的结构来表示边不相交路径的连接图，这种结构是一个二维表的形式，节点表示行，路径表示列。另外还提出了双射和(bijective sum)合成关系，接合(splice)合成关系以及合成关系的图实现操作，双射和是用来表示连接2个具有k条不相交路径的子图后形成的k+1条不相交路径的图，这个图包括1个具有k–1条不相交路径的公共子图和添加到共享节点(指在2个具有k条路径的图中公共子图连接另外一条路径的节点)上的2条附加路径。接合是把一个图中属于2个不同路径上的2个节点合并成一个节点的过程，以此确保挖掘的完全性。合成关系的图实现是对于给定的一个有n条路径的合并关系C，则可以构造一个相对应的图。让表中的行对应图的顶点并且定义边(i，j)为：当同一个路径p的2个节点出现在第i行和第j行中，则在它们之间的那条边。这种操作就称为图实现。

算法的处理主要分成3个阶段：

（1）通过每次增加一条边来构造频繁路径；

（2）构造路径数为2的频繁图。通过连接具有一条路径的图来完成，在此过程通过合成关系列出所有的两条路径连接的可能性，另外保留那些在图实现中路径数为2、而且通过计算支持度确定的频繁图，最后移除所有的非最小的同构图；

（3）通过连接具有k–1条路径的图来构造具有k条路径的频繁图。使用双射和方法，把找到的2个具有k条路径的图(它们具有相同的k–1条路径的公共子结构)连接成一个k+1条路径的图模式。但是因为使用双射和直接合成2个图模式可能会造成路径上个别的公共顶点的丢失，因此必须使用接合方法来弥补这一缺陷，添加丢失的公共顶点给候选模式以确保完全性。接下来移除同构图并计算支持度来确定频繁性。最后移除所有不是最大的频繁子图。

1.2 基于模式增长的频繁子结构挖掘算法

当连接2个大小为k的频繁子结构产生大小为k+1的图候选时，基于Apriori算法的系统开销很大，为避免这种系统开销，提出了模式增长的方法，主要包括gSpan、CloseGraph和FFSM等。这些算法均通过逐步扩展频繁边得到频繁子图，但每个算法对图的扩展过程也有许多不同之处。

1.2.1 gSpan 算法

gSpan 算法[5]旨在减少复制图二度发现的图)的产生。它首次提出利用 DFS(深度优先搜索)法生成频繁子图，通过两大技术的应用—— DFS 词典序、最小 DFS 编码和最右扩展，对每个图建立 DFS 词典序，并达到将每个图用最小 DFS 编码唯一标记的目的，使得无需按 Apriori 算法的思想而直接生成频繁子图。该算法通过选择一个起始顶点开始访问，并为能分辨出已经访问过的顶点对其做标记，然后对被访问过的顶点集合反复扩展，直到建立一个完全的深度优先搜索(DFS)树。在构造 DFS 树时，顶点的访问顺序形成一个线性序(用下标来记录此次序)，设起始顶点为根，则最后访问的顶点称为最右顶点，从根到最右顶点的直接路径称为最右路径。gSpan 扩展时只进行最右扩展，即在 DFS 树中一条新边可以添加到最右顶点和最右路径上另一个顶点之间或者引进一个新的顶点并连接到最右路径上的顶点。把每个加下标的图转换为边序列称为 DFS 编码，用{i, j, li, $l(i, j)$, lj}5 元组表示，然后通过一定规则来建立边序列之间的序，即 DFS 词典序，基于词典序，找到图的最小 DFS 编码。只有对最小 DFS 编码执行最右扩展，才能减少复制图的产生，也确保了挖掘结果的完全性。gSpan 无论在计算时间上还是内存消耗上都是一种高效的方法。但是由于它对图模式的表示有一个非常严格的顺序，于是有人又提出了挖掘闭频繁图的 CloseGraph 算法。

1.2.2 CloseGraph 算法

CloseGraph 算法[6]提出了一些新的方法，如同等出现(Equivalent Occurrence)和提前终止(Early Termination)，利用这些方法可以大大减少没必要的子图的生成，最终提高挖掘效率。

给定图 g 和图数据集 D={G_1, G_2, …, G_n}，设 $\tau(g, D)$为 g 在 D 的每个图中的子图同构的总数目，图 g 可以通过增加一个新边 e 来扩展形成新的图 g'，令 $\zeta(g, g', D)$为在 D 的每个图中 g(对应于 g')的可扩展的子图同构的总数目。如果 $\tau(g, D) = \zeta(g, g', D)$成立，则称 g 和 g'同等出现，即意味着在 D 中 g 出现时 g'一定出现。如果 $g \subset g'$并且 $g' \not\subset g''$ 能推出 g'' 不是闭的，此时仅需要扩展 g'来代替扩展 g，这种情况称为提前终止。

CloseGraph 算法的执行主要分 3 个步骤：

(1)生成一个频繁图；

(2)根据一个频繁图 g 是闭的当且仅当不存在与 g 具有相同支持度的真超图 g'，亦即如果想知道一个图是否是闭的，仅需要检查比它多一条边的超图的支持度。如果二者支持度不相等，则 g 是闭的，否则不是闭的。通过这条规则可以检查(1)中生成的图是否是闭频繁图；

(3)检查提前终止的条件和任何一种可能导致提前终止失败的情况，来决定此生成图是否应该被扩展。

CloseGraph 算法不仅能够减少不必要的生成子图而且也能充分地提高挖掘的效率，特别是在挖掘大型图数据集时(比如说多于 32 条边的较大的频繁图)，它的性能大概可以优于 gSpan 性能的 4 ~ 10 个因子。

1.2.3 FFSM 算法

FFSM 算法[7]采用深度逐层递归来挖掘频繁子图。每个图均用一个标准邻接矩阵 CAM(Canonical Adjacency Matrix)来描述，它使用一种独特的表示图结构的规范形式并且提出两种有效的候选操作：FFSM 联接操作(子图“交”)和 FFSM 扩展操作；使用一种代数图结构(非最佳标准的 CAM 树)能够完全地列举出所有的频繁子图；能通过对每 1 个频繁子图的嵌入集合测试完全避免子图同构。其中矩阵的联接操作是合并两个矩阵形成一个矩阵集，而对一个矩阵 M 的扩展操作也会产生一个矩阵集，集合中的每一个矩阵增加了一个额外的节点以及连接此节点和 M 中最后一个节点的一条边。

1.3 基于最小描述长度的近似频繁子结构挖掘算法

SUBDUE[8]是一个基于图的学习系统，该系统的输入可以是带标记或不带标记的简单图或图集，采用了最小描述长度(MDL)原则，挖掘近似的频繁子结构，并将它们精确地表示出来。它根据最小描述长度原则，输出最好的压缩输入集的子结构模式，它采用了约束搜索(beam search)方法，通过扩展节点递增地增长单个顶点。每次扩展它都搜索最佳总描述长度：模式的描述长度和图集合的描述长度，模式全部实例都浓缩成单个节点。在发现了最好的子结构以后输入图就被重写，下一次迭代使用重写的图作为一个新的输入图，这样，在每次迭代中，算法仅能找到一个子结构。此外 SUBDUE 进行近似匹配，允许子结构有轻微变化，从而支持近似子结构的发现，而且它也能以预先确定子图的形式潜入到背景知识里去。

1.4 基于模式增长和模式归约的精确稠密频繁子结构挖掘算法

关系图是一种特殊的图结构，其中每个节点标号在每个图中仅用一次，它被广泛地应用于大型网络(例如生物网络、社会网络、交通网络和万维网)的建模和分析中。在大型关系图中频繁高连通的子图或稠密子图是一种令人感兴趣的模式。这种模式有助于在社会

网络中识别具有紧密联系的人群，高连通的子图也可以在生物学中表示相同功能组件中基因的集合。

CloseCut 和 Splat 就是用来在大型关系图中(大约有 10 k 个节点和 1M 条边的关系图)挖掘具有连通性约束的闭频繁图模式，采用边连通性的概念并运用相关的图论知识来加速挖掘过程。

CloseCut 实际上是一种基于模式增长的方法，它首先开始于一个小的频繁候选图，通过增加新边来尽可能地扩展，直到找到具有相同支持度的闭超图，在闭超图中把先前候选图的顶点压缩成为一个顶点。然后分解这个高连通性的闭超图，判断每个顶点的度是否满足连通性约束条件，提取满足连通性约束的子图，删除所有的与不满足连通性约束的顶点连接的边。然后，通过增加新边来扩展候选图，并且重复进行上述操作直到没有候选图是频繁的。

Splat 是一种模式归约的方法，它代替从小到大的枚举图，而且直接对关系图取交并分解它们来得到高连通图。令模式 g 是关系图 $Gi1$，$Gi2$，…，$Gil(i1<i2<\cdots<il)$的高连通图。为了在更大的集合$\{Gi1, Gi2, \cdots, Gil, Gil+1\}$中挖掘模式，取 $g'=g \cap Gil+1$。G 中的一些边可能被删除，因为它们不在图 $Gil+1$ 中，因此，新图 g' 的连通性可能不再满足约束，这需要将 g' 分解成较小的高连通子图，通过求交和分解操作逐渐地减小候选图的大小。最终，它可能变成零图。这就是模式归约的方法。

2 现有频繁子结构挖掘算法的分析比较

上述方法均是基于图论的数据挖掘方法。基于 Apriori 的方法必须使用宽度优先搜索(BFS)策略，因为它逐层产生候选。这种方法为了确定大小为 $k+1$ 的图是否频繁，必须检查它的所有对应的大小为 k 的子图来获得其频度的上界。这样，在挖掘任何大小为 $k+1$ 的子图之前，类 Apriori 的方法通常必须完成大小为 k 的子图的挖掘。因此类 Apriori 的算法需要采用 BFS，相反，模式增长方法在搜索方式上更加灵活一些，它既可以使用宽度优先搜索，也可以使用深度优先搜索(DFS)，它要比基于 Apriori 的方法占用较少的内存[4]。

AGM 和 FSG 算法都利用邻接矩阵分别对图的顶点和边进行逐层构造，以最终获取频繁子图。所不同的是，AGM 求出了导出子图，图不一定连通，而 FSG 则以边为每次迭代的对象，求出了连通的频繁子图。

gSpan 算法比 AGM、FSG 算法计算更高效，而且它采取从内存到磁盘交换数据，减少了内存的消耗。gSpan 和 FSG 算法能够找到所有符合用户要求的子图，但是不可否认的是它们都产生大量的子结构，而 subDue 的特色就在于它能高效地发现较少数量的但更有趣的最好地压缩子结构模式。而这些可能是 gSpan 和 FSG 不能发现的，但是在发现子图同构方面 Subdue 不比 gSpan 和 FSG 具有更高的效率，还需要进一步扩展。

CloseGraph 算法类似于 gSpan，但是它只挖掘闭频繁子图而且在对一个已经生成的图进行最右扩展之前，先检查该图是否存在提前终止。这样 CloseGraph 可以经常产生更少的图模式，因此比挖掘全部模式集合的 gSpan 更有效。

FFSM 算法能够回避图与图之间直接的同构测试，通过使用一种代数图方法能高效地处理子图同构的基本问题，它的性能优于 gSpan 算法。特别是对合成数据集使用 FFSM 算法更高效。

稠密子结构的两种挖掘算法在大的图数据集上具有良好的可伸缩性。在对具有高支持度和低连通性的模式，CloseCut 具有更好的性能，相反，在挖掘的初期阶段，Splat 能够过滤低连通性的频繁图，对于高连通性约束，它具有更好的性能。但是，当关系图的数量增加时，它需要枚举出取并的关系图，此时 Splat 性能可能会恶化。

在频繁子图挖掘算法中，需要解决的问题是子图同构问题和找出所有频繁子图的方法上，从频繁子图的挖掘上，已经取得了较大进展，上述这些算法都能够在满足一定的要求下，找到所需要的结果，但是子图同构问题仍没得到很好地解决(子图同构问题已被证明是 NP 完全问题)，需要对算法进一步扩展，有待进一步研究。

总结本文所提到的各种频繁子结构挖掘算法的共同特征及各自特点如表 1 所示。

图挖掘已经具有了广泛的应用，包括化学、生物学、材料科学、通讯网络等领域。利用图挖掘方法挖掘出的各种频繁子结构可以被应用在大型图数据库中建立图索引、进行结构相似性搜索，对结构数据集合的特征化以及进行图数据集的分类和聚类分析。虽然图挖掘领域的研究刚刚起步，但是因为图拓扑结构在数学方面是最重要的结构之一，并且同逻辑语言密切相关，因此图挖掘理论和技术将会在数据挖掘和机器学习方面作出突出的贡献，并将在各个领域得到广泛的实际应用。

参考文献

[1] INOKUCHI A, WASHIO T, MMOTODA H.An apriori-based algorithm for mining frequent substructures from graph data.In Proc. 2000 European Symp.Principle of Data Mining and Knowledge Discovery(PKDD'00),Lyon,France,1998(9):13-33.

表1 频繁子结构挖掘典型算法的总结比较

类型	典型算法及其主要应用特点		共同特征
基于Apriori的方法	AGM	在图集中挖掘子图模式，用于合成密集型数据集中性能更优	逐层产生候选，采用宽度优先搜索技术，产生大量频繁子图，系统开销很大
	FSG	在图集中挖掘频繁连通子图模式，特别适用于大型图数据集	
基于模式增长的方法	gSpan	在图集中挖掘频繁连通子图模式，FFSM特别适用于合成数据集	既可采用宽度优先搜索也可以采用深度优先搜索，占用较少内存，与类Apriori的方法相比普遍调高了挖掘效率提高
	FFSM		
	closeGraph：挖掘频繁闭合连通子图模式，特别适用于大型图数据集，能有效地去除冗余的频繁子图，减少结果集大小，并能保证不丢失任何信息		
基于MDL原则的近似子图挖掘方法	subDue：针对近似模糊的生物化合物结构挖掘，适用于超大规模稀疏图，进行近似匹配允许少许结构变化，能高效地发现较少数量的但更有趣的最好地压缩子结构模式		利用启发式搜索策略，使用模糊计算方法，减少了挖掘时间，提高了挖掘效率，但不能保证结果集的完整性
稠密了图挖掘方法	CloseCut Splat	在大型关系图中挖掘具有连通性约束的闭频繁稠密子结构模式	在大的图数据集上具有良好的可伸缩性

[2] KURAMOCHI M, KARYPIS G.Frequent subgraphs discovery.In Proc.2001 Int.Conf.Data Mining(ICDM'01),San Jose,CA,2001(11):313-320.

[3] VANETIK N, GUDES E, SHIMONY S.E.Computing frequent graph patterns from semistructured data.In Proc.2002 Int.Conf.on Data Mining(ICDM'02),Maebashi,Japan,2002(10):458-465.

[4] HAN J, KAMBER M.数据挖掘概念与技术[M].北京:机械工业出版社, 2007.

[5] YAN X.HAN J.gSpan:Graph-based substructure pattern mining. In Proc.2002 Int.Conf.Data Mining(ICDM'02),Maebashi,Japan,2002(10):721-724.

[6] YAN X, HAN J.Closegraph:mining closed frequent graph patterns. In KDD,2003:286-295.

[7] HUAN J, WANG W, PRINS J.Efficient mining of frequent subgraphs in the presence of isomorphism.In ICDM, 2003:549-552.

[8] KETKAR S, HOLDER B, COOK J.Subdue:Compression-Based Frequent Pattern Discovery.In ACM, 2005:71-76.

（收稿日期：2009-02-01）

(上接第4页)

Computers and Electronics in Agriculture, 1998,20:117-130.

[12]MOLTO E, BLASCO J, BENLLOCH J V. Computer vision for automatic inspection of agricultural produces. In SPIE Symposium on Precision Agriculture and Biological Quality, 1 - 6 November, 1998, Boston, MA, USA.

[13]SHAHIN M A, TOLLNER E W, MCCLENDON R W. Apple classification based on surface bruises using image processing and neural networks[J]. Transactions of the ASAE, 2002,45(5),1619-1627.

[14] LEEMANS V, DESTAIN M F. A real-time grading method of apples based on features extracted from defects[J]. Journal of Food Engineering, 2004(61):83-89.

[15]何东健,杨 青.果实缺陷面积的计算机视觉测定研究[J].农业工程学报,1997(12):156-160.

[16]刘 禾,汪懋华.用计算机图像技术进行苹果坏损自动检测的研究[J].农业机械学报,1998(12):81-86.

[17]王江枫,罗锡文.计算机视觉技术在芒果重量及果面坏损检测中的应用.农业工程学报,1998(12):186-189.

[18]应义斌,景寒松,马俊福,等.机器视觉技术在黄花梨尺寸和果面缺陷检测中的应用.农业工程学报,1999(1).

[19]邓继忠,张泰岭.应用计算机视觉技术对梨碰压伤的检测[J].农业工程学报,1999(3):205-209.

[20]何东健,耿 楠.用活动边界模型精确检测果实表面缺陷[J].农业工程学报,2001(9):159-162.

[21]冯斌,汪懋华.计算机视觉技术识别水果缺陷的一种新方法[J].中国农业大学学报,2002(4):73.

[22]王亚琴,高华,自然环境下水果分割与定位研究[J].计算机工程,2004(7):583-586.

[23] RUIZ L A, MOLTO E, JUSTE F et al. Location and characterization of the stem calyx area on oranges by computer vision[J]. Journal of Agricultural Engineering Research, 1996,(64): 165-172.

[24] PENMAN D W. Determination of stem and calyx location on apples using automatic visual inspection[J]. Computers and Electronics in Agriculture, 2002,(33): 7-18.

[25] KAVDIR I B, GUYER D E. Comparison of artificial neural networks and statistical classifiers in apple sorting using textural features [J]. Biosystems Engineering, 2004,89 (3), 331-344.

[26] BlASCO J, ALEIXOS N, MOLTO E. Machine vision system for Automatic quality grading of fruit[J]. Biosystems Engineering, 2003, 85 (4), 415-423.

(收稿日期：2009-02-25)