ISSN 1000-9825, CODEN RUXUEW
Journal of Software
doi: 10.3724/SP.J.1001.2008.03473

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

从不确定图中挖掘频繁子图模式*

邹兆年, 李建中+, 高 宏, 张 硕

(哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

Mining Frequent Subgraph Patterns from Uncertain Graphs

ZOU Zhao-Nian, LI Jian-Zhong+, GAO Hong, ZHANG Shuo

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: E-mail: lijzh@hit.edu.cn, http://db.hit.edu.cn

Zou ZN, Li JZ, Gao H, Zhang S. Mining frequent subgraph patterns from uncertain graphs. *Journal of Software*, 2008. http://www.jos.org.cn/1000-9825/3473.htm

Abstract: This paper studies uncertain graph data mining and especially investigates the problem of mining frequent subgraph patterns from uncertain graph data. A data model is introduced for representing uncertainties in graphs, and an expected support is employed to evaluate the significance of subgraph patterns. By using the apriori property of expected support, a depth-first search-based mining algorithm is proposed with an efficient method for computing expected supports and a technique for pruning search space, which reduces the number of subgraph isomorphism testings needed by computing expected support from the exponential scale to the linear scale. Experimental results show that the proposed algorithm is 3 to 5 orders of magnitude faster than a naïve depth-first search algorithm, and is efficient and scalable.

Key words: uncertain graph; graph mining; frequent subgraph pattern

摘 要: 研究不确定图数据的挖掘,主要解决不确定图数据的频繁子图模式挖掘问题.介绍了一种数据模型来表示图的不确定性,以及一种期望支持度来评价子图模式的重要性.利用期望支持度的 Apriori 性质,给出了一种基于深度优先搜索策略的挖掘算法.该算法使用高效的期望支持度计算方法和搜索空间裁剪技术,使得计算子图模式的期望支持度所需的子图同构测试的数量从指数级降低到线性级.实验结果表明,该算法比简单的深度优先搜索算法快 3~5 个数量级,有很高的效率和可扩展性.

关键词: 不确定图;图挖掘;频繁子图模式

中图法分类号: TP311 **文献标识码**: A

近年来,科研领域积累了大量用图来建模的数据(简称图数据),如化合物分子结构、传感器网络拓扑结构、社会网络等.从图数据中发现有用的知识已成为一项重要研究课题,称为图挖掘.现已提出大量图挖掘算法[1−7].这些算法全部针对确定图数据(即完整且精确的图数据).然而,在许多实际应用中还存在大量的不确定图数据,

* Supported by the National Natural Science Foundation of China under Grant Nos.60533110, 60773063 (国家自然科学基金) the National Basic Research Program of China under Grant No.2006CB303005 (国家重点基础研究发展计划(973)); the Program for New Century Excellent Talents in University of China under Grant No.NCET-05-0333 (新世纪优秀人才支持计划)

Received 2008-05-29; Accepted 2008-10-09; Published online 2008-12-22

例如,生物信息学中的蛋白质交互(protein-protein interaction,简称 PPI)网络是一类不确定图,其顶点表示蛋白质,边表示蛋白质交互.由于 PPI 实验检测方法的局限性,很大一部分检测到的 PPI 是不确定的.文献[8]提出一种 PPI 可靠性指数来衡量 PPI 真实存在的可能性.图 1 给出了文献[8]中的一个 PPI 网络实例,其中,顶点上的文字表示蛋白质的功能,边上的整数表示 PPI 的可靠性指数.可靠性指数越小,表示 PPI 真实存在的可能性越大.

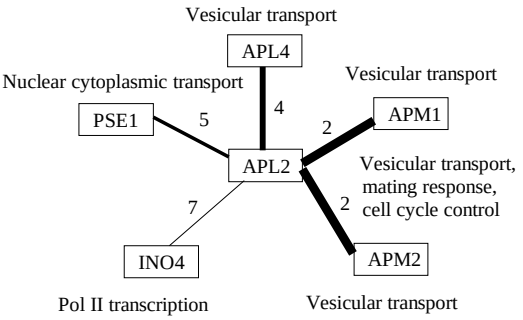


Fig.1 An example of PPI network

图 1 蛋白质交互网络实例

挖掘不确定图数据具有十分重要的实际意义.例如,生物学家通常对蛋白质功能之间联系的结构感兴趣.通过挖掘实验得到的 PPI 网络,生物学家可以很方便地获得有关蛋白质功能间联系的结构的知识.但如上所述,这类图数据是不确定的.因此,在这些图数据中,不确定性成为评价知识的重要性的一项重要指标.只有当某种蛋白质功能之间联系的结构频繁出现且出现的可能性很大时,该结构才被视为有用的知识,而出现可能性小的结构一般是没有实际用处的.因此,挖掘不确定图数据在实际应用中具有重要的应用价值.

本文中,不确定图是一种边带有权值的特殊加权图,边的权值表示该边在其两个端点之间实际存在的可能性.不确定图数据库是一个不确定图的集合.不确定图与确定图的不同之处在于,一个不确定图代表了由它蕴含的全部确定图上的概率分布.确定图 I 被不确定图 G 蕴含,若 I 和 G 具有相同的顶点集且 I 的边集是 G 的边集的子集.一个不确定图数据库则代表了由它蕴含的全部确定图数据库上的概率分布.确定图数据库 $d=\{I_1,I_2,\ldots,I_n\}$ 被不确定图数据库 $D=\{G_1,G_2,\ldots,G_n\}$ 蕴含,若对所有 $1\le i\le n$ 有 G_i 蕴含 I_i.第 2 节将给出不确定图数据模型的详细定义.

本文研究从不确定图数据库中挖掘频繁子图模式.在传统的频繁子图模式挖掘中,一个子图模式在确定图数据库中出现的比例称为该子图模式的支持度.然而,此定义在不确定图数据中没有意义,因为一个子图模式是否能够匹配到一个不确定图中是不确定的.实际上,子图模式 S 在不确定图数据库 D 中的支持度是 S 在所有 D 蕴含的确定图数据库中的支持度上的概率分布.为了便于处理,我们使用这个概率分布的期望值来度量子图模式的重要性.该期望值称为子图模式的期望支持度.若一个子图模式的期望支持度大于等于给定的阈值 *minsup*,则该子图模式是频繁的.因此,本文研究的问题可以描述为:给定不确定图数据库 D 和阈值 *minsup*,找出 D 中全部频繁子图模式.

不确定图数据上的频繁子图模式挖掘问题向我们提出了如下挑战:

(1) 不确定图数据库的子图模式的数量是不确定图数量的指数.枚举全部子图模式是不现实的.因此,需要一种高效的方法,能够从全部子图模式中找出全部频繁子图模式.

(2) 计算子图模式在不确定图数据库 D 中的期望支持度需要计算 D 蕴含的全部确定图数据库上的概率分布.由于 D 蕴含的确定图数据库的数量非常大,这种计算期望支持度的方法显然是不可行的.因此,需要一种计算期望支持度的优化方法.

针对上述挑战,本文提出了一种挖掘不确定图数据库中的全部频繁子图模式的算法.本文的主要贡献为:

(1) 证明子图模式的期望支持度满足 Apriori 性质,即一个频繁子图模式的任何子图都是频繁的.并利用 Apriori 性质对全部子图模式构成的搜索空间进行深度优先搜索.

(2) 提出一种计算子图模式期望支持度的高效方法.该方法不需要计算不确定图数据库蕴含的全部确定图数据库上的概率分布,并将计算所需的子图同构测试数量从指数级降低到线性级.

(3) 提出一种有效的子图模式搜索空间裁剪技术.该技术可以有效降低对子图模式搜索空间进行深度优先搜索的开销.

(4) 进行了大量的实验来考察算法的效率、可扩展性以及边的不确定性对算法性能的影响.

本文第 1 节介绍本文的相关工作.第 2 节给出数据模型并定义不确定图数据上的频繁子图模式挖掘问题.第 3 节给出本文的算法及理论分析.第 4 节给出实验结果.第 5 节总结本文的工作.

1 相关工作

传统的频繁子图模式挖掘都是在确定图数据上进行的.文献[1,2]分别提出基于广度优先搜索的 AGM 算法和 FSG 算法.文献[3–5]分别提出基于深度优先搜索的 gSpan 算法、FFSM 算法和 Gaston 算法.为了减少冗余的子图模式,文献[6]提出了 CloseGraph 算法挖掘闭合频繁子图模式.文献[7]提出 SPIN 算法挖掘极大频繁子图模式.尽管目前已经存在许多频繁子图模式挖掘算法,但所有这些算法都无法应用到不确定图数据的挖掘中.

不确定数据管理方面的研究工作主要包括不确定数据的建模[9]、关系查询[10]、Top-k 查询[11]以及 Skyline 查询[12].然而,不确定数据挖掘仍然是一项挑战.在不确定数据的频繁项集挖掘方面,文献[13]提出一种基于 Apriori 性质的 U-Apriori 算法和裁剪技术.文献[14]改进了文献[13]中的裁剪技术.文献[15]中提出了在不确定数据流上挖掘频繁项的精确算法及随机近似算法.然而,这些算法都无法扩展到不确定图数据上.

2 问题定义

定义 1. 不确定图是一个四元组 $G=((V,E),\Sigma,L,P)$,其中(V,E)是一个无向图,V 是顶点集,E 是边集,Σ是标记集,$L:V\cup E\to\Sigma$是顶点和边的标记函数, $P:E\to(0,1]$ 是边的存在可能性函数.

边的存在可能性表示边在其两个端点之间实际存在的可能性.1 表示边一定存在.从而,确定图就是一个所有边的存在可能性皆为 1 的特殊的不确定图,可记为一个三元组$((V,E),\Sigma,L)$.一个不确定图实际蕴含着一组确定图.确定图 $I=((V',E'),\Sigma',L')$被不确定图 $G=((V,E),\Sigma,L,P)$所蕴含(记作 $G\Rightarrow I$),若 $V'=V,E'\subseteq E,\Sigma'=\Sigma,L'=L|_{V'\cup E'}$,其中 $L|_{V'\cup E'}$表示将 L 约束在 $V'\cup E'$上得到的函数.为使模型简单,本文假定不确定图中不同的边存在与否是相互独立的.因此,不确定图 $G=((V,E),\Sigma,L,P)$蕴含确定图 $I=((V',E'),\Sigma',L')$的可能性为

$$P(G\Rightarrow I)=\prod_{e\in E'}P(e)\cdot\prod_{e\in E-E'}(1-P(e)) \tag{1}$$

公式(1)成立是因为所有 E'中的边都出现在 I 中,并且所有 $E–E'$中的边都不出现在 I 中.令 $Imp(G)$表示 G 蕴含的所有确定图的集合.由于 G 中边的不同组合,$Imp(G)$中包含 $2^{|E|}$个确定图.我们有如下重要的定理.

定理 1. 对于一个不确定图 G,函数 $P(G\Rightarrow I)$定义了样本空间 $Imp(G)$上的一个概率分布.

不确定图数据库是一个不确定图的集合.一个不确定图数据库实际上蕴含着一组确定图数据库.确定图数据库 $d=\{I_i|1\le i\le n\}$被不确定图数据库 $D=\{G_i|1\le i\le n\}$蕴含(记作 $D\Rightarrow d$),若对于所有 $1\le i\le n$ 有 $G_i\Rightarrow I_i$.令 $Imp(D)$表示 D 蕴含的所有确定图数据库的集合.显然,$Imp(D)$中包含 $\prod_{i=1}^{n}2^{|E_i|}$ 个确定图数据库.假定不确定图数据库中的不确定图是相互独立的,则不确定图数据库 D 蕴含确定图数据库 d 的可能性为

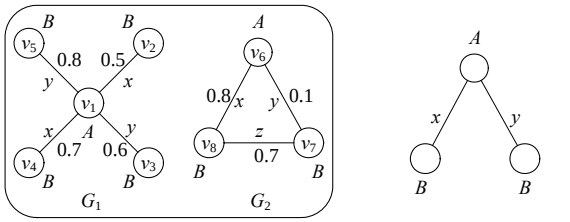
$$P(D\Rightarrow d)=\prod_{i=1}^{n}P(G_i\Rightarrow I_i) \tag{2}$$

定理 2. 对于一个不确定图数据库 D,函数 $P(D\Rightarrow d)$定义了样本空间 $Imp(D)$上的一个概率分布.

例 1:图 2(a)给出一个不确定图数据库 $D=\{G_1,G_2\}$,其中顶点上的文字表示顶点的标记,例如 G_1 中的顶点 v_1 的标记为 A;边上的文字表示边的标记,例如 G_1 中的边(v_1,v_2)的标记为 x;边上的数字表示边的存在可能性,例如 G_1 中的边(v_1,v_2)的存在可能性为 0.5.因此,G_1 表示它蕴含的 2^4=16 个确定图上的概率分布(如图 3 所示).G_2 表示它蕴含的 2^3=8 个确定图上的概率分布.D 表示它蕴含的 16×8=128 个确定图数据库上的概率分布.

定义 2. 确定图 $G=((V,E),\Sigma,L)$子图同构于确定图 $G'=((V',E'),\Sigma',L')$(记作 $G\subseteq_C G'$),当且仅当存在一个单射

$f:V\to V'$满足:(1) $\forall v\in V,L(v)=L'(f(v))$;(2) $\forall(u,v)\in E,(f(u),f(v))\in E'$;(3) $\forall(u,v)\in E,L((u,v))=L'((f(u),f(v)))$.$G'$的子图$(V'',E'')$称为在子图同构$f$下$G$在$G'$中的嵌入,其中,$V''=\{f(v)|v\in V\}$,$E''=\{(f(u),f(v))|(u,v)\in E\}$.



(a) Uncertain graph database D (b) Subgraph pattern S

(a) 不确定图数据库 D (b) 子图模式 S

Fig.2 An example of uncertain graph database

图 2 不确定图数据库示例

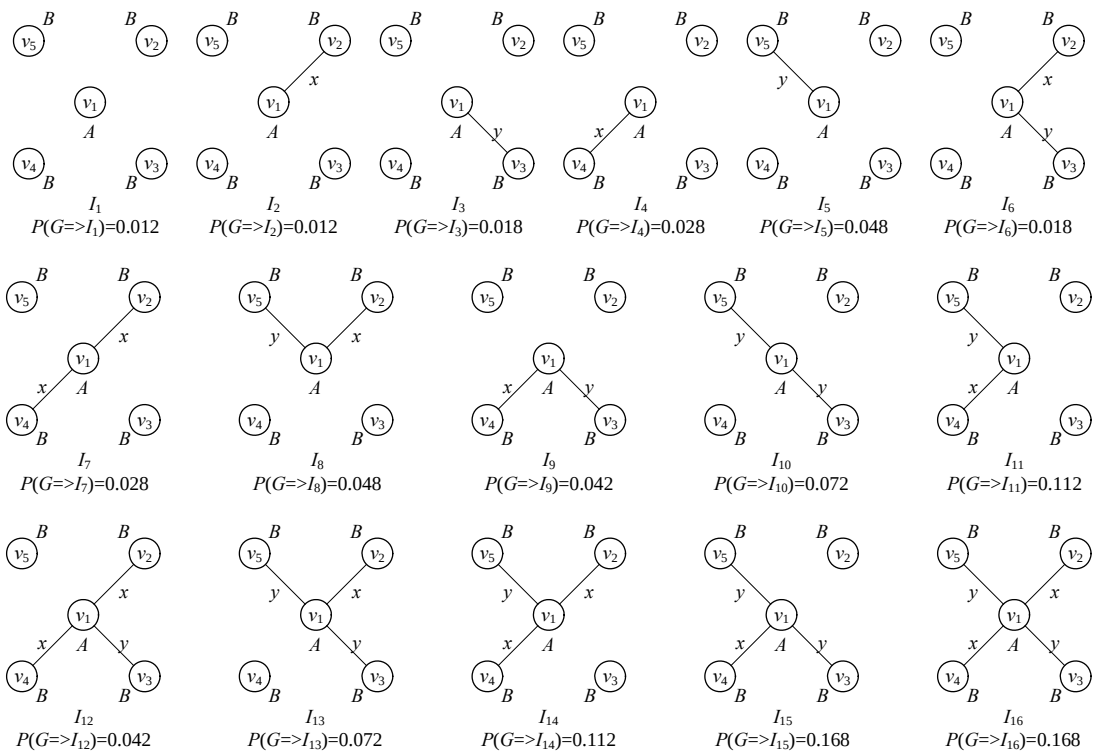


Fig.3 Probability distribution of all certain graphs implicated by uncertain graph G_1 in Fig.2

图 3 图 2 中的不确定图 G_1 蕴含的全部确定图集合上的概率分布

在传统频繁子图模式挖掘中,子图模式是一个连通子图,它与确定图数据库中至少 1 个确定图子图同构.子图模式 S 在确定图数据库 D 中的支持度为 $sup_D(S)=|\{G|S\subseteq_C G,G\in D\}|/|D|$.然而,这些概念在不确定图数据库中没有意义,因为一个确定子图是否包含在一个不确定图中是不确定的.因此,需要重新定义这些概念.

定义 3. 连通的确定图 S 是不确定图数据库 D 的子图模式,若 S 子图同构于 D 中至少 1 个不确定图 G 蕴含的某个确定图.如果 S 是 D 的子图模式且边数为 k,则 S 是 D 的一个 k-子图模式.对于两个子图模式 S 和 S',若 $S\subseteq_C S'$,则 S 是 S'的子模式,S'是 S 的超模式.S 是 S'的直接子模式当且仅当 S 是 S'的子模式且$|E_S|+1=|E_{S'}|$.

定义 4. 对于不确定图数据库 D,设 $Imp(D)$为 D 蕴含的所有确定图数据库的集合,则子图模式 S 在 D 中的

支持度是一个概率分布:

$$\begin{bmatrix} s_1 & s_2 & \dots & s_m \\ P(s_1) & P(s_2) & \dots & P(S_m) \end{bmatrix},$$

其中,$s_1,s_2,\dots,s_m$ 是 S 在 $Imp(D)$中的确定图数据库中的传统支持度,$P(s_i)=\sum_{d\in Imp(D)|sup_d(S)=s_i} P(D\Rightarrow d)$ 是支持度 s_i 的概率,$m=|\{sup_d(S)|d\in Imp(D)\}|$.

定义 5. 设子图模式 S 在不确定图数据库 D 中的支持度为定义 4 中的概率分布,则 S 在 D 中的期望支持度为

$$esup_D(S)=\sum_{i=1}^{m} s_i\cdot P(s_i)=\sum_{d\in Imp(D)} sup_d(S)\cdot P(D\Rightarrow d) \tag{3}$$

子图模式 S 在不确定图数据库 D 中是频繁的,若 S 在 D 中的期望支持度不小于给定的阈值 $minsup\in[0,1]$.因此,不确定图数据库上的频繁子图模式挖掘问题可以定义为:给定不确定图数据库 D 和阈值 $minsup$,找出 D 中全部频繁子图模式的集合 $FP=\{S|S$ 是 D 的子图模式,且 $esup_D(S)\geq minsup\}$.

子图模式 S 出现在不确定图 G 中(记作 $S\subseteq_U G$),若 S 子图同构于至少 1 个 G 蕴含的确定图.因此,S 在 G 中出现的概率为

$$P(S\subseteq_U G)=\sum_{I\in Imp(G)} P(G\Rightarrow I)\cdot\psi(I,S) \tag{4}$$

其中,$\psi(I,S)=1$,若 $S\subseteq_C I$;否则,$\psi(I,S)=0$.因此,利用公式(4)可将公式(3)变为

$$esup_D(S)=\frac{1}{|D|}\cdot\sum_{i=1}^{|D|} P(S\subseteq_U G_i) \tag{5}$$

本文使用公式(5)计算 S 在 D 中的期望支持度,因为公式(5)中求和项的数量$|D|$远远小于公式(3)中求和项的数量$|Imp(D)|$.由公式(4)和公式(5),可得如下重要结论:

(1) 给定不确定图 G,子图模式在 G 中出现的概率满足 Apriori 性质,即对于任意子图模式 S 和 S',若 S 是 S' 的子模式,则 $P(S\subseteq_U G)\geq P(S'\subseteq_U G)$;

(2) 给定不确定图数据库 D,子图模式在 D 中的期望支持度满足 Apriori 性质,即对于任意子图模式 S 和 S',若 S 是 S'的子模式,则 $esup_D(S)\geq esup_D(S')$.

根据期望支持度的 Apriori 性质,频繁子图模式的任何子模式都是频繁的,不频繁子图模式的任何超模式都是不频繁的.该性质可被用来降低频繁子图模式挖掘算法的复杂度.

3 频繁子图模式挖掘算法

3.1 算法概述

给定输入不确定图数据库 D 和阈值 $minsup$.D 中子图模式间的直接子模式关系(见定义 3)构成一个偏序关系,因此 D 中所有子图模式可以被组织成一个格,这个格可用一个有向无环图来表示,称为 D 的子图模式搜索空间.D 中全部 k-子图模式均被安排在搜索空间的第 k 层.图 4 给出了图 2 中的不确定图数据库 D 的子图模式搜索空间,其中节点代表子图模式,边代表子图模式之间的直接子模式关系.假定搜索空间第 1 层上的 1-子图模式分别包含边 $e_1,e_2,\dots,e_k$,则搜索空间可被划分为 k 个互不相交的子搜索空间,其中,第 i 个子搜索空间中的任何子图模式都包含边 e_i,但都不包含边 $e_1,e_2,\dots,e_{i-1}$.图 4 中的子搜索空间被多边形框起.

将 D 中所有子图模式组织成一个搜索空间后,频繁子图模式挖掘问题转化为如何高效地遍历搜索空间来枚举全部频繁子图模式的问题.本文的算法采用深度优先搜索策略依次对每个子搜索空间进行深度优先搜索来枚举全部频繁子图模式.其工作过程如下:

对于每个子搜索空间,深度优先搜索从该子搜索空间第 1 层的 1-子图模式开始.对于深度优先搜索过程中访问到的每个子图模式 S,首先计算 S 在 D 中的期望支持度 $esup_D(S)$,若 $esup_D(S)\geq minsup$,即 S 是频繁的,则将 S 加入结果集并继续深度优先搜索 S 的后裔(即 S 的全部超模式);若 $esup_D(S)<minsup$,即 S 是非频繁的,则由 Apriori 性质,S 的全部超模式也是非频繁的,因此停止对 S 的后裔进行深度优先搜索并回溯到最后一个在深度优先搜索

中已被访问过的 S 的父亲节点.当深度优先搜索回溯至该子搜索空间第 1 层的 1-子图模式且该子图模式的所有儿子节点都已被访问时,深度优先搜索停止.此时,该子搜索空间中的全部频繁子图模式均已遍历完毕.

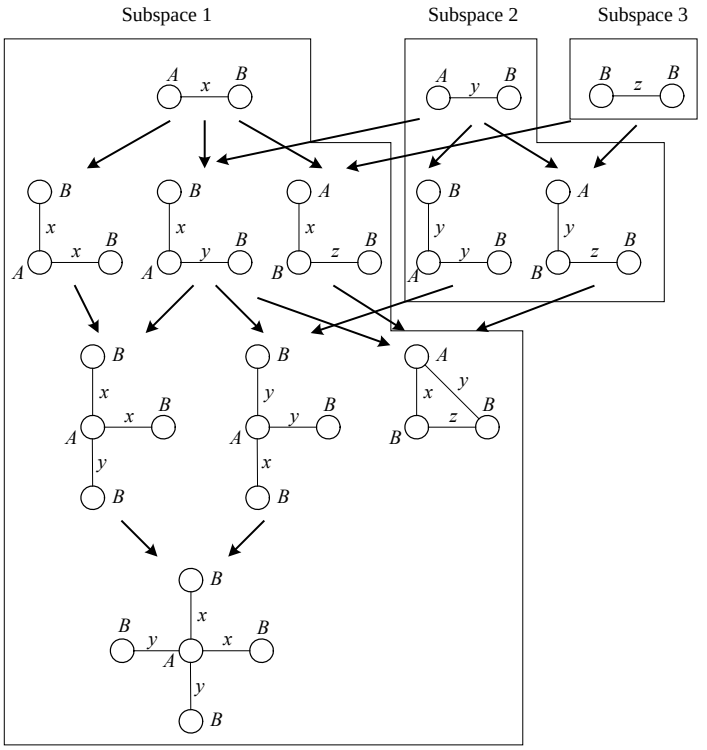


Fig.4 Search space of subgraph patterns

图 4 子图模式搜索空间

值得注意的是,由于搜索空间是一个有向无环图,因此搜索空间中的一个节点 S 可以有多个父亲节点,从而在深度优先搜索过程中 S 可能会被多次访问.为避免对 S 的后裔进行重复深度优先搜索,算法需要判断 S 是否已被访问过,若 S 已被访问过,则停止对 S 的后裔进行深度优先搜索.

显然,本文算法的关键在于期望支持度的计算和搜索空间的剪裁.第 3.2 节将给出一种计算期望支持度的高效方法.第 3.3 节将介绍一种有效的子图模式空间裁剪技术.第 3.4 节将给出完整的算法.

3.2 计算期望支持度的高效算法

根据公式(5),计算子图模式 S 在不确定图数据库 D 中的期望支持度 $esup_D(S)$的主要困难在于计算 S 在 D 中每个不确定图 G_i 中出现的概率 $P(S\subseteq_U G_i)$.根据公式(4),计算 $P(S\subseteq_U G_i)$需要计算 G_i 蕴含的全部 $2^{|E_{G_i}|}$ 个确定图上的概率分布,并进行从 S 到 G_i 蕴含的每个确定图的子图同构测试.因此,这种简单的计算方法总共需要进行 $\sum_{i=1}^{|D|}2^{|E_{G_i}|}$ 次子图同构测试.为了降低计算复杂度,本节提出一种计算子图模式期望支持度的高效算法.该算法在计算 S 在 D 中的期望支持度时只需进行$|D|$次子图同构测试.

3.2.1 算法描述

设 $G\Rightarrow\Gamma(S_1,S_2,\ldots,S_k)$表示如下事件:不确定图 G 蕴含一组确定图$\Gamma(S_1,S_2,\ldots,S_k)$,其中每个确定图都包含 S_1,S_2,…,S_k 作为其子图.由公式(1)可得事件 $G\Rightarrow\Gamma(S_1,S_2,\ldots,S_k)$发生的概率为

$$P(G\Rightarrow\Gamma(S_1,S_2,...,S_k))=\sum_{I\in\Gamma(S_1,S_2,...,S_k)}P(G\Rightarrow I)=\prod_{e\in E_{S_1}\cup E_{S_2}\cup...\cup E_{S_k}}P(e) \quad (6)$$

计算子图模式的期望支持度的高效算法基于下面两个重要的定理.

定理 3. 给定不确定图 G 和子图模式 S.设 S 在 G 中嵌入的集合为$\{S_1,S_2,\ldots,S_l\}$,则 S 在 G 中出现的概率为

$$P(S\subseteq_U G)=\sum_{1\le i\le \ell}P(G\Rightarrow \Gamma(S_i))-\sum_{1\le i_1<i_2\le \ell}P(G\Rightarrow \Gamma(S_{i_1},S_{i_2}))+\ldots+$$
$$(-1)^{j-1}\sum_{1\le i_1<i_2<\ldots<i_j\le \ell}P(G\Rightarrow \Gamma(S_{i_1},S_{i_2},\ldots,S_{i_j}))+\ldots+(-1)^{\ell-1}\sum_{1\le i_1<i_2<\ldots<i_\ell\le \ell}P(G\Rightarrow \Gamma(S_{i_1},S_{i_2},\ldots,S_{i_\ell})) \quad (7)$$

定理 4. 设 $S_1,S_2,\ldots,S_k$ 是子图模式 S 在不确定图 G 中的 k 个嵌入.若其中 S_p 与 S_q 边不相交(1≤p<q≤k),即 $E_{S_p}\cap E_{S_q}=\varnothing$, 则

$$P(G\Rightarrow\Gamma(S_1,S_2,\ldots,S_k))=P_1\cdot P_2/P_3 \quad (8)$$

其中,
$$P_1=P(G\Rightarrow\Gamma(S_1,S_2,\ldots,S_{p-1},S_{p+1},\ldots,S_k)),$$
$$P_2=P(G\Rightarrow\Gamma(S_1,S_2,\ldots,S_{q-1},S_{q+1},\ldots,S_k)),$$
$$P_3=P(G\Rightarrow\Gamma(S_1,S_2,\ldots,S_{p-1},S_{p+1},\ldots,S_{q-1},S_{q+1},\ldots,S_k)).$$

根据定理 3 和定理 4,我们首先给出计算子图模式 S 在不确定图 G 中出现的概率 $P(S\subseteq_U G)$的算法.算法中,为快速检查两个嵌入是否边不相交,算法使用一个称为嵌入图的数据结构.具体来说,给定 S 在 G 中的嵌入的集合 EM,嵌入图是一个图 $EG=(V_{EG},E_{EG})$,其中顶点集 V_{EG} 表示 EM 中的全部嵌入,边集 E_{EG} 表示 EM 中的嵌入之间的边相交关系,即对于任意 $S_i,S_j\in V_{EG}$ 有$(S_i,S_j)\in E_{EG}$ 当且仅当 $E_{S_i}\cap E_{S_j}\neq\varnothing$. 显然,$EM$ 中两个嵌入 S_i 与 S_j 边不相交当且仅当$(S_i,S_j)\notin E_{EG}$.算法的过程如下:

算法 1. COMP-OCC-PROB.

输入:子图模式 S 和不确定图 G.

输出:S 在 G 中出现的概率 $P(S\subseteq_U G)$.

步骤 1. 计算 S 在 G 中的全部嵌入的集合 EM 并构造嵌入图 EG.

步骤 2. 按 k 从 1 到$|EM|$的递增顺序,计算 EM 的全部包含 k 个元素的子集$\{S_1,S_2,\ldots,S_k\}$的 $P(G\Rightarrow\Gamma(S_1,S_2,\ldots,S_k))$.

(1) 当 k=1 时,对任意包含 1 个元素的子集$\{S_i\}\subseteq EM$,按公式(6)计算 $P(G\Rightarrow\Gamma(S_i))$.

(2) 当 k=2 时,对任意包含 2 个元素的子集$\{S_i,S_j\}\subseteq EM$,若 S_i 与 S_j 在 EG 中没有边相连,即 S_i 与 S_j 边不相交,则 $P(G\Rightarrow\Gamma(S_i,S_j))=P(G\Rightarrow\Gamma(S_i))\cdot P(G\Rightarrow\Gamma(S_j))$,否则,按公式(6)计算 $P(G\Rightarrow\Gamma(S_i,S_j))$.

(3) 当 3≤k≤$|EM|$时,对任意包含 k 个元素的子集$\{S_1,S_2,\ldots,S_k\}\subseteq EM$,若存在边不相交的嵌入 $S_p,S_q\in\{S_1,S_2,\ldots,S_k\}$,则按公式(8)计算 $P(G\Rightarrow\Gamma(S_1,S_2,\ldots,S_k))$,否则,按公式(6)计算 $P(G\Rightarrow\Gamma(S_1,S_2,\ldots,S_k))$.

步骤 3. 按公式(7)计算 S 在 G 中出现的概率 $P(S\subseteq_U G)$并输出 $P(S\subseteq_U G)$.

利用算法 1,我们给出计算子图模式期望支持度的算法.算法的过程如下.

算法 2. COMP-EXP-SUP.

输入:子图模式 S 和不确定图数据库 D.

输出:S 在 D 中的期望支持度 $esup_D(S)$.

步骤 1. 对 D 中每个不确定图 G_i,使用算法 1 计算 S 在 G_i 中出现的概率 $P(S\subseteq_U G_i)$.

步骤 2. 按公式(5)计算 S 在 D 中的期望支持度 $esup_D(S)$并输出 $esup_D(S)$.

3.2.2 算法分析

由定理 3、定理 4 以及公式(5)可得算法 1 和算法 2 是正确的.

在算法 2 中,对于 D 中每个不确定图 G_i,算法 2 使用算法 1 计算 S 在 G_i 中出现的概率,其中只需一次子图同构测试得到 S 在 G_i 中的全部嵌入.因此,算法 2 总共需要进行的子图同构测试次数为$|D|$.这比简单的期望支持度计算方法所需的 $\sum_{i=1}^{|D|}2^{|E_{Gi}|}$ 次子图同构测试少得多,即从指数数量级降到线性数量级.

引理 1. 给定子图模式 S 在不确定图 G 中的嵌入集合的任意子集$\{S_1,S_2,\ldots,S_k\}$.若存在边不相交的嵌入 $S_p,S_q\in\{S_1,S_2,\ldots,S_k\}$,则算法 1 必然在 $O(1)$时间内计算 $P(G\Rightarrow\Gamma(S_1,S_2,\ldots,S_k))$.

定理 5. 给定不确定图 G 和子图模式 S.设 EG 是描述 S 在 G 中的嵌入集合的嵌入图.若 EG 包含 n 个顶点

和 m 条边,则公式(7)中可以在 $O(1)$时间内计算的项 $P(G \Rightarrow \Gamma(S_{i_1}, S_{i_2}, ..., S_{i_j}))$ 所占的比例至少为 $1-(2^n-1)^{-1}(2^{n'}+2^{m'}+n-n'-2)$.其中,$n'$和 m'是整数,满足 $\binom{n'}{2} \le m, \binom{n'+1}{2} > m, \binom{n'}{2} + m' = m.$

图 5 给出了当 EG 的顶点数 n=10 时,公式(7)中所有可以在 $O(1)$时间内计算的项所占的比例的下界与 EG 的边数 m 的关系.图 5 说明,该比例的下界是很高的.例如,当 EG 包含 m=20 条边时,该比例至少为 $1-(2^{10}-1)^{-1}\times(2^6+2^5+10-5-2)\approx 90\%$.由此可见,算法 1 在实际应用中非常有效.

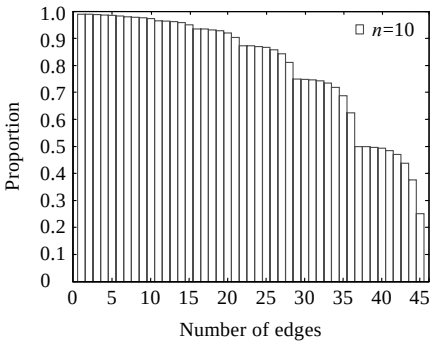


Fig.5 Relationship between the proportion of terms in Eq.(7) that can be computed in $O(1)$ time and the number of edges in EG

图 5 公式(7)中可在 $O(1)$时间内计算的项的比例与 EG 的边数的关系

在实际应用中,由于图的稀疏性,子图模式在图中的嵌入有很大的可能不重叠,即嵌入图 EG 远不是一个完全图,即 m 远远小于 $\binom{n}{2}$,且 m 和 n 在同一数量级上.由于 $\binom{n'}{2} \le m$,可知 $n' = O(\sqrt{n})$,并且由于 $n<m$,有 $2^{n'} = 2^{O(\sqrt{n})} > n$.因此,该比例至少为 $O(1-2^{-\sqrt{n}})$,这说明公式(7)中大部分的项都可以在 $O(1)$时间内计算.

3.3 子图模式搜索空间裁剪技术

在挖掘算法的深度优先搜索过程中,当被访问的子图模式 S 的期望支持度 $esup_D(S)$小于 $minsup$ 时,由 Apriori 性质,S 及其全部超模式都是非频繁的,因此可以从搜索空间中裁剪掉.由算法 2 可知,为了计算 S 的期望支持度,必须计算 S 在每个不确定图 G_i 中出现的概率 $P(S \subseteq_U G_i)$.为了降低挖掘算法的开销,本节提出一种有效的搜索空间裁剪技术.该技术不需要计算全部 $P(S \subseteq_U G_i)$即可裁剪不频繁子图模式 S.该方法基于如下定理:

定理 6. 给定不确定图数据库 D.设 S 和 S'是任意两个子图模式且 S'是 S 的子模式,则对于 k=0,1,…,$|D|$,

$$esup_D(S) \le \frac{1}{|D|}\left(\sum_{i=1}^{k} P(S \subseteq_U G_i) + \sum_{i=k+1}^{|D|} P(S' \subseteq_U G_i)\right) \tag{9}$$

不等式(9)右边的表达式是 S 的期望支持度的上界.若不等式(9)的右边小于 $minsup$,则 S 一定是非频繁的.因此,可以通过计算不等式(9)的右边而非精确的 $esup_D(S)$来决定 S 是否为非频繁的.该方法能够减少计算开销,因为不等式(9)中全部 $P(S' \subseteq_U G_i)$已经在访问 S 之前得到,我们只需对 i=1,2,…,k 计算 $P(S \subseteq_U G_i)$.

结合算法 2,我们得到计算子图模式的期望支持度上界的算法.算法的过程如下:

算法 3. COMP-UPP-SUP.

输入:子图模式 S、不确定图数据库 D 和阈值 $minsup$.

输出:S 在 D 中的期望支持度 $esup_D(S)$的上界.

步骤 1. 初始化 k=1,$upper$=0.S'=在深度优先搜索中最后一个已被访问过的 S 的父亲节点.

步骤 2. 若 $k>|D|$,则输出 $upper$,算法结束;否则,使用算法 1 计算 S 在 G_k 中出现的概率 $P(S \subseteq_U G_k)$,并计算不等式(9)右边的表达式,得到 $esup_D(S)$的上界并赋值到 $upper$.

步骤 3. 若 $upper<minsup$,则输出 $upper$,算法结束;否则,$k=k+1$ 并转到步骤 2 继续执行.

值得注意的是,若子图模式 S 是频繁的,则算法 3 在步骤 2 结束.根据公式(5),此时 $upper=esup_D(S)$.

3.4 完整算法

把第 3.2 节和第 3.3 节的优化算法加入第 3.1 节的算法框架,得到完整的频繁子图模式挖掘算法.

算法 4. MINE-FREQ-SUBG.

输入:不确定图数据库 D 和阈值 $minsup$.

输出:D 中全部频繁子图模式的集合 FP.

1. 初始化 FP 为空集;
2. 扫描 D 得到 D 中全部 1-子图模式;
3. **FOR** D 中每个 1-子图模式 S **DO**
4. 调用子过程 $DFS(S,D)$;
5. 输出 FP;

算法 4 子过程. $DFS(S,D)$.

6. 使用算法 3 计算 S 在 D 中的期望支持度 $esup_D(S)$的上界 $upper$;
7. **IF** $upper<minsup$ **THEN** 子过程结束;
8. $FP=FP\cup\{S\}$;
9. 扫描 D 得到 S 的全部直接超模式;
10. **FOR** S 的每个直接超模式 S' **DO**
11. **IF** S'尚未被访问过 **THEN** 调用子过程 $DFS(S',D)$.

算法 4 如下工作:初始时,结果集 FP 为空.算法扫描 D 中每条边得到 D 中全部 1-子图模式.然后对每个 1-子图模式 S,算法调用递归子过程 DFS 对以 S 为根的子搜索空间进行深度优先搜索,发现其中的频繁子图模式.最后,算法输出结果集 FP.

子过程 DFS 的功能是对子搜索空间进行深度优先搜索来发现频繁子图模式.其过程如下:首先使用算法 3 计算 S 在 D 中的期望支持度 $esup_D(S)$的上界 $upper$.若 $upper<minsup$,则子过程返回.此时 S 是非频繁的,停止对 S 及其所有超模式进行深度优先搜索.否则,将 S 加入到 FP 中,扫描 D 得到 S 的全部直接超模式.然后对 S 的每个直接超模式 S',若 S'尚未被访问过,则递归调用子过程 $DFS(S',D)$,对 S'的所有超模式进行深度优先搜索,发现其中的频繁子图模式.为了降低检查子图模式是否已被访问过的复杂性,可将子图模式进行编码.若一个子图模式的编码不是其标准编码,则该子图模式必然已被访问过.许多编码方式(如 DFS 编码[3]和 CAM 编码[4])可以在这里使用.

4 实 验

我们进行了大量的实验来考察本文算法的执行效率、可扩展性以及不确定性对算法效率的影响.算法使用 C 语言实现.用于实验的计算机具有 Intel Core2 Duo 2GHz CPU 和 2GB 内存,运行 Windows XP 操作系统.

实验数据集按如下步骤产生:

步骤 1. 使用文献[2]中的图数据生成器生成一个确定图数据集.该数据生成器有 6 个输入参数:D(图的数量),V(顶点标记的数量),E(边标记的数量),I(频繁子图模式的平均大小),L(潜在频繁子图模式数量),T(图的平均大小).

步骤 2. 为步骤 1 生成的确定图数据集中的每条边赋予一个存在可能性.存在可能性服从均值为 m、方差为 d^2 的正态分布.

在实验 1 中,我们实现了一种只利用 Apriori 性质的深度优先搜索算法 NAÏVE,并分别在 6 个数据集和 4 种阈值 $minsup$ 下比较了本文算法和 NAÏVE 算法的执行时间.由于一个不确定图蕴含的全部确定图的数量是该不确定图边数的指数,因此 NAÏVE 算法只能挖掘包含几十个不确定图的数据集.实验结果见表 1.表 1 中的"–"表示 NAÏVE 算法因内存耗尽而终止.其中,数据集 D_1 的参数为 D=20,L=10,V=5,E=1,I=5,T=10,m=0.9,d=0.1;数据集

D_2 的参数为 D=40,L=10,V=5,E=1,I=5,T=10,m=0.9,d=0.1;数据集 D_3 的参数为 D=20,L=10,V=5,E=1,I=5,T=15,m=0.9,d=0.1;数据集 D_4 的参数为 D=20,L=10,V=5,E=1,I=5,T=10,m=0.8,d=0.1;数据集 D_5 的参数为 D=20,L=10,V=1,E=10,I=5,T=10,m=0.9,d=0.1;数据集 D_6 的参数为 D=20,L=10,V=5,E=1,I=7,T=10,m=0.9,d=0.1.实验结果说明,本文的算法比 NAÏVE 算法快 3~5 个数量级.因此,本文提出的期望支持度计算方法和子图模式搜索空间裁剪技术非常有效.

Table 1 Execution time comparison between the proposed algorithm and the NAÏVE algorithm

表 1 本文算法与 NAÏVE 算法执行时间的比较

Datasets	*minsup* (%)	Execution time (s)		Datasets	*minsup* (%)	Execution time (s)	
		Our algorithm	Algorithm NAÏVE			Our algorithm	Algorithm NAÏVE
D_1	80	0.00	0.86	D_4	80	0.00	0.93
	40	0.00	5.80		40	0.00	4.00
	20	0.01	55.71		20	0.01	32.13
	10	0.03	171.61		10	0.02	123.05
D_2	80	0.00	3.20	D_5	80	0.00	0.22
	40	0.01	9.42		40	0.08	9.29
	20	0.01	65.57		20	0.25	36.45
	10	0.08	232.25		10	3.26	114.27
D_3	80	0.01	15.93	D_6	80	0.00	443.64
	40	0.01	53.48		40	0.03	3 865.90
	20	0.42	311.05		20	0.15	–
	10	0.53	2 635.54		10	0.45	–

在实验 2 中,我们考察在阈值 *minsup* 变化的情况下本文算法的执行时间.实验分别在 8 个数据集 D_7,…,D_{14} 上进行.D_7 的参数为 D=20000,L=100,V=10,E=10,I=5,T=20,m=0.95,d=0.05.D_8 的参数为 D=10000,L=100,V=10,E=10,I=5,T=30,m=0.95,d=0.05.D_9 的参数为 D=10000,L=200,V=10,E=10,I=5,T=20,m=0.95,d=0.05.D_{10} 的参数为 D=10000,L=100,V=5,E=10,I=5,T=20,m=0.95,d=0.05.D_{11} 的参数为 D=10000,L=100,V=10,E=5,I=5,T=20,m=0.95,d=0.05.D_{12} 的参数为 D=10000,L=100,V=10,E=10,I=10,T=20,m=0.95,d=0.05.D_{13} 的参数为 D=10000,L=100,V=10,E=10,I=5,T=20,m=0.8,d=0.05.D_{14} 的参数为 D=10000,L=100,V=10,E=10,I=5,T=20,m=0.95,d=0.01.实验结果如图 6(a)所示.实验结果说明,本文算法的执行时间随 *minsup* 的增加而显著减少.这是因为随着 *minsup* 的增加,频繁子图模式的数量显著减少(如图 6(b)所示),从而本文算法需要进行的子图同构测试和支持度计算的数量都将显著下降,故算法的执行时间减少.

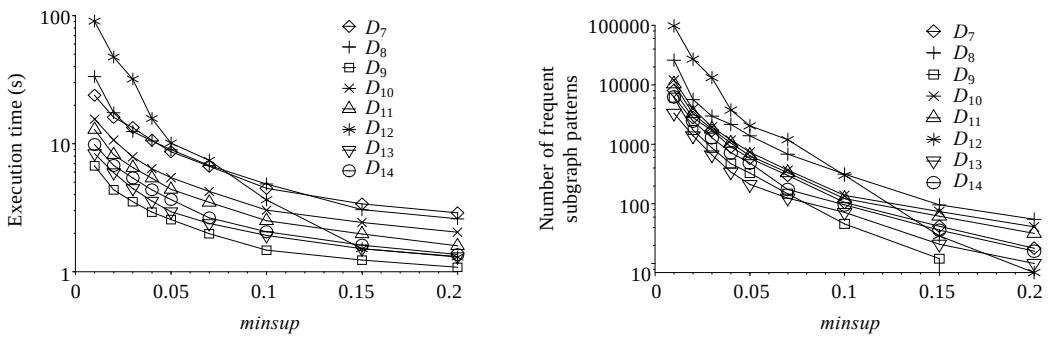


Fig.6 Impact of the variation of *minsup* on the execution time of the algorithm

图 6 *minsup* 的变化对算法执行时间的影响

在实验 3 中,我们考察在不确定图的数量增加的情况下本文算法的可扩展性.衡量可扩展性的指标是算法执行时间和子图同构测试次数.实验分别在实验 2 中使用过的 7 个数据集 D_8,…,D_{14} 上进行.实验中,阈值 *minsup* 取 10%.实验结果如图 7 所示.实验结果说明,本文算法的执行时间和子图同构测试次数随不确定图的数量的增

加而呈线性增长.因此,本文算法具有非常高的可扩展性.

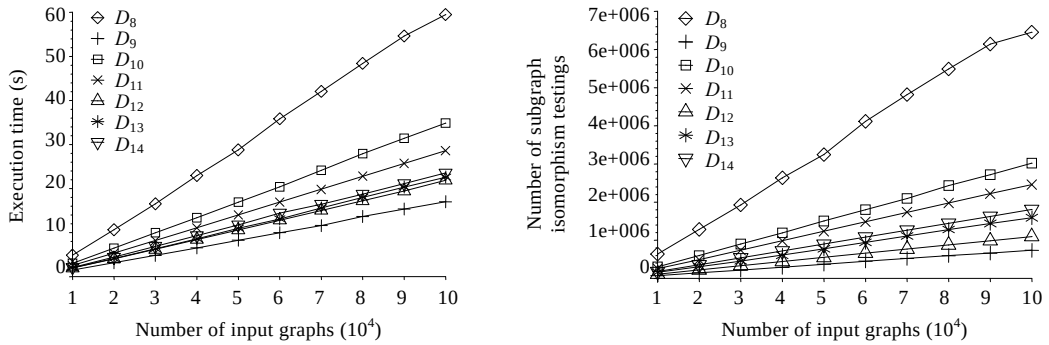


Fig.7 Impact of the variation of the number of input uncertain graphs on the scalability of the algorithm

图 7 不确定子图数量的变化对算法可扩展性的影响

在实验 4 中,我们首先考察在不确定图中边存在可能性的均值变化的情况下本文算法的执行时间.实验分别在 4 个数据集上进行.数据集的参数为 D=10000,L=100,V=10,E=10,I=5,T=40,d 分别取 0.01,0.05,0.1 和 0.2.实验中,阈值 *minsup* 取 1%.实验结果如图 8(a)所示.实验结果说明,随着边存在可能性的均值的增加,本文算法的执行时间也增加.这是因为随着边存在可能性的均值的增加,子图模式在不确定图中存在的概率提高,故期望支持度增大,频繁子图模式的数量增加,因此本文算法的执行时间也相应增加.

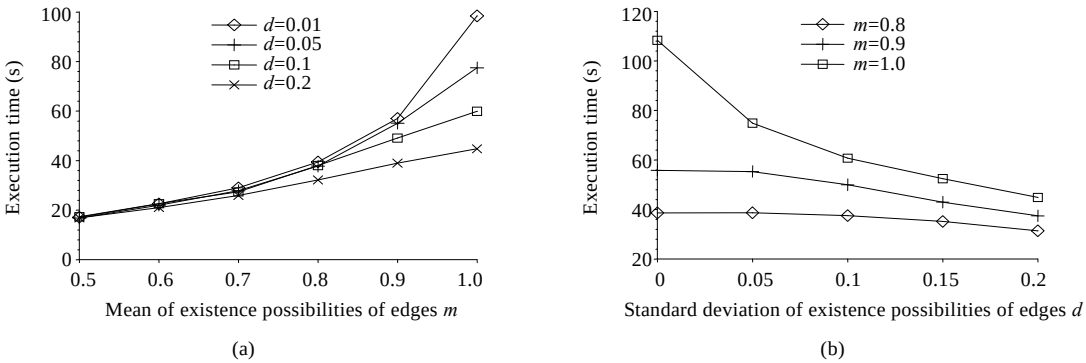


Fig.8 Impact of the variation of existence possibilities of edges on the execution time of the algorithm

图 8 边存在可能性的变化时算法的执行时间的影响

在实验 4 中,我们还考察在不确定图中边的存在可能性的方差变化的情况下本文算法的执行时间.实验分别在 3 个数据集上进行.数据集的参数为 D=10000,L=100,V=10,E=10,I=5,T=40,m 分别取 0.8,0.9 和 1.0.实验中,阈值 *minsup* 取 1%.实验结果如图 8(b)所示.实验结果说明,随着边存在可能性的方差的增加,本文算法的执行时间显著减少.这是因为随着边存在可能性的方差的增加,具有较低存在可能性的边的数量也随之增加,从而导致频繁子图模式数量的减少,因此本文算法的执行时间也减少.

5 结 论

本文提出了在不确定图数据库中挖掘频繁子图模式的问题,并给出一种解决该问题的高效算法.由于期望支持度满足 Apriori 性质,本文算法采用对子图模式搜索空间的深度优先搜索策略.影响算法性能的关键因素是子图模式期望支持度的计算和子图模式搜索空间的裁剪.本文提出的期望支持度计算方法将所需的子图同构测试次数从指数级降低到线性级.本文提出的子图模式空间裁剪技术显著提高了裁剪子图模式搜索空间的效

率.实验结果表明,本文算法远远优于简单的深度优先搜索方法,并且有很高的效率和线性的可扩展性.

References:

[1] Inokuchi A, Washio T, Motoda H. An apriori-based algorithm for mining frequent substructures from graph data. In: Zighed DA, Komorowski HJ, Zytkow JM, eds. Proc. of the 4th European Conf. on Principles of Data Mining and Knowledge Discovery. Lyon: Springer-Verlag, 2000. 13−23.

[2] Kuramochi M, Karypis G. Frequent subgraph discovery. In: Cercone N, Lin TY, Wu X, eds. Proc. of the 2001 IEEE Int'l Conf. on Data Mining. San Jose: IEEE Computer Society, 2001. 313−320.

[3] Yan X, Han J. gSpan: Graph-Based substructure pattern mining. In: Kumar V, Tsumoto S, Zhong N, Yu PS, Wu X, eds. Proc. of the 2002 IEEE Int'l Conf. on Data Mining. Maebashi: IEEE Computer Society, 2002. 721−724.

[4] Huan J, Wang W, Prins J. Efficient mining of frequent subgraphs in the presence of isomorphism. In: Kumar V, Tsumoto S, Zhong N, Yu PS, Wu X, eds. Proc. of the 2003 IEEE Int'l Conf. on Data Mining. Melbourne: IEEE Computer Society, 2002. 549−552.

[5] Nijssen S, Kok JN. A quickstart in frequent structure mining can make a difference. In: Kim W, Kohavi R, Gehrke J, DuMouchel W, eds. Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Seattle: ACM, 2004. 647−652.

[6] Yan X, Han J. Closegraph: Mining closed frequent graph patterns. In: Getoor L, Senator TE, Domingos P, Faloutsos C, eds. Proc. of the 9th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Washington: ACM, 2003. 286−295.

[7] Huan J, Wang W, Prins J, Yang J. Spin: Mining maximal frequent subgraphs from graph databases. In: Kim W, Kohavi R, Gehrke J, DuMouchel W, eds. Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Seattle: ACM, 2004. 581−586.

[8] Saito R, Suzuki H, Hayashizaki Y. Interaction generality, a measurement to assess the reliability of a protein-protein interaction. Nucleic Acids Research, 2002,30(5):1163−1168.

[9] Benjelloun O, Sarma AD, Halevy A, Theobald M, Widom J. Databases with uncertainty and lineage. VLDB Journal, 2008,17(2): 243−264.

[10] Dalvi N, Suciu D. Efficient query evaluation on probabilistic databases. VLDB Journal, 2007,16(4):523−544.

[11] Soliman MA, Ilyas IF, Chang KCC. Top-*k* query processing in uncertain databases. In: Chirkova R, Dogac A, Ozsu T, Sellis T, eds. Proc. of the 2002 IEEE Int'l Conf. on Data Engineering. Istanbul: IEEE Computer Society, 2007. 896−905.

[12] Pei J, Jiang B, Lin X, Yuan Y. Probabilistic skylines on uncertain data. In: Koch C, Gehrke J, Garofalakis MN, Srivastava D, Aberer K, Deshpande A, Florescu D, Chan CY, Ganti V, Kanne CC, Klas W, Neuhold EJ, eds. Proc. of the 24th Int'l Conf. on Very Large Data Bases. Austria: ACM, 2007. 15−26.

[13] Chui CK, Kao B, Hung E. Mining frequent itemsets from uncertain data. In: Zhou Z, Li H, Yang Q, eds. Proc. of the 11th Pacific-Asia Conf. on Advances in Knowledge Discovery and Data Mining. Nanjing: Springer-Verlag, 2007. 47−58.

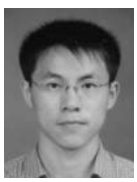
[14] Chui CK, Kao B. A decremental approach for mining frequent itemsets from uncertain data. In: Washio T, Suzuki E, Ting KM, Inokuchi I, eds. Proc. of the 12th Pacific-Asia Conf. on Advances in Knowledge Discovery and Data Mining. Osaka: Springer-Verlag, 2008. 64−75.

[15] Zhang Q, Li F, Yi K. Finding frequent items in probabilistic data. In: Wang JT, ed. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. Vancouver: ACM, 2008. 819−832.

邹兆年(1979－),男,吉林长春人,博士生,主要研究领域为图数据挖掘.

李建中(1950－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据库系统,传感器网络.

高宏(1966－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据仓库,传感器网络.



张硕(1982－),男,博士生,主要研究领域为图数据管理.