第33卷 第8期 计 算 机 学 报 Vol. 33 No. 8
2010年8月 CHINESE JOURNAL OF COMPUTERS Aug. 2010

RAKING：一种高效的不确定图 K-极大频繁模式挖掘算法

韩 蒙[1] 张 炜[2] 李建中[1),2)]

[1)](黑龙江大学计算机科学技术学院 哈尔滨 150080)

[2)](哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

摘 要 由于不确定图蕴含了指数级的可能图实例，基于确定图模型的频繁图模式挖掘算法通常难以在不确定图集合上高效运行.文中提出了一种不确定图数据集上的基于随机游走的 K 极大频繁子模式挖掘算法.首先，将每个不确定图转换为相应的确定图并挖掘候选频繁模式；然后，将候选频繁模式恢复为不确定图并生成极大频繁模式搜索空间；最后，通过随机游走以相同概率随机地选择 K 个极大频繁模式.理论分析和实验结果表明文中提出的算法能够高效地获得不确定图集合的 K-极大频繁模式.

关键词 不确定图；数据挖掘；随机游走；极大频繁模式

中图法分类号 TP18 **DOI号**：10.3724/SP.J.1016.2010.01387

RAKING: An Efficient K-Maximal Frequent Pattern Mining Algorithm on Uncertain Graph Database

HAN Meng[1)] ZHANG Wei[2)] LI Jian-Zhong[1),2)]

[1)](*School of Computer Science and Technology, Heilongjiang University, Harbin* 150080)

[2)](*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001)

Abstract An uncertain graph can represent a large number of possible graph instances. This greatly reduces the efficiency of existing frequent pattern mining algorithms. The paper proposes a random walk based K-maximal frequent pattern mining algorithm on uncertain graph set. Firstly, each uncertain graph is converted to a graph without uncertain information. Candidate frequent patterns are retrieved from the converted graph set. Then, the candidate frequent patterns are transformed to corresponding uncertain graph pattern and searching space of maximal frequent patterns are constructed as well. Finally, K-maximal frequent patterns are selected from all maximal frequent patterns equiprobably. Theoretical analysis and experimental results show that the proposed algorithm can efficiently retrieve the K-maximal frequent patterns of an uncertain graph set.

Keywords uncertain graph; data mining; random walk; maximal frequent pattern

1 引 言

图模型被广泛应用于描述化学信息学、生物信息学、医学和社会科学等领域的科学数据.如何高效地实现图数据集上的数据挖掘已成为数据库研究领域的热点问题之一.在实际应用中，很多图数据是以不确定的形式存在的.例如，生物信息学中，由于对

收稿日期：2010-06-11.本课题得到国家自然科学基金(60903017)和黑龙江大学学生学术科技创新项目(2010183,2010204,2010208)资金资助.韩 蒙，男，1987年生，硕士研究生，主要研究方向为图数据挖掘.E-mail：hanmeng-2005@163.com.张 炜，男，1975年生，博士，讲师，主要研究方向为图数据管理、无线传感器网络.李建中，男，1950年生，教授，博士生导师，主要研究领域为并行数据库、传感器网络.

基因及蛋白质的测量和实验手段存在着人为或客观的误差,使一些分子结构或片段的属性无法确定.同时,研究对象的结构及组成有时也会不断发生动态变化.因此,在这些以图模型描述的数据对象中,结点和边的属性值通常在一定值域内满足某种概率分布,而边的存在性也通过概率进行表达.在社会网络中,人与人的关系往往存在一定的不确定性;另外,在化学及医学数据中也都存在着类似的情况.因为不确定性在日常生产和生活中广泛存在,所以针对不确定图集合的数据挖掘方法的研究具有十分重要的意义.然而,目前这类研究工作还非常少.

虽然目前已提出很多基于确定图模型的数据挖掘方法[1-12],但这些方法通常无法直接挖掘不确定图数据集.设 n 为不确定图的结点和边的总数.基于可能世界模型,每个不确定图蕴含 2^n 个可能的图实例.于是,整个不确定图数据集合蕴含了指数级的可能图实例.以极大频繁模式挖掘为例,目前在确定图上提出的极大频繁模式的挖掘方法通常都采用"两步法"的方式:首先,获得全部频繁集;然后,在频繁集上通过剪枝或二次挖掘获得所需的极大频繁模式集.然而,在不确定图集合上直接使用这类算法则需要枚举和遍历整个指数级的可能图实例空间来进行图模式的分析,这无疑是十分低效的.另外,由于可能图实例数量巨大,相应的频繁模式数目也会急剧增加而难以分析.因此,如何选择输出合适的频繁图模式,也是基于不确定图的频繁模式挖掘必须要解决的问题.

针对不确定图集合上频繁模式挖掘所面临的挑战,本文提出了一种基于随机游走的 K-极大频繁模式挖掘算法 RAKING(Random wAlk based K-maxImal uNcertain frequent patterns mininG algorithm).

不确定图上的极大频繁模式具有以下优点:首先,一个不确定极大频繁模式蕴含了所有可能频繁模式的不确定信息;其次,K-极大频繁模式虽然只是频繁模式的一个小子集,但是它却代表了整个频繁集的重要特征.不确定图数据集中的极大频繁模式挖掘在科研和实际应用中具有十分重要的作用.例如,分析蛋白质交互作用网络时,通过对极大频繁模式的挖掘,生物学家可以很方便地获得出现最频繁且最可能发生的各个子功能团的极大集.很多研究工作只需在这个极大集上进行处理即可,不再需要对数量巨大的频繁集进行复杂的分析.同时,极大频繁模式也蕴含了各个可能存在的蛋白质功能团的有关信息.

RAKING 算法主要包括以下 3 个步骤:

1. 挖掘候选频繁子图模式.将不确定图集合中的每个不确定图转换为一个对应的确定图.然后,在转换后的确定图集合上使用挖掘算法获取候选频繁子图模式.接下来再将所获得的频繁子图模式恢复为不确定图.在预处理过程中,该算法避免枚举指数级的可能图实例,从而有效地减小了搜索空间.

2. 生成随机游走空间.首先,随机选择一个预处理后的候选频繁模式.然后,以这个模式为起点,构建带有权值的模式增长空间,形成合适的随机游走空间.

3. 获得 K-极大频繁模式.在搜索空间上进行随机游走,依据 Apriori 性质有效剪枝并判断当前模式是否为极大频繁模式.若是则输出该模式,否则继续随机游走.

需要注意的是,RAKING 算法只输出 K 个极大频繁模式.若不确定图数据集的极大频繁模式多于 K 个,则算法在极大频繁模式集中以相同概率随机选择 K 个极大频繁模式输出.于是,RAKING 算法既提供了输出模式数量的调节机制,也给出了输出模式的选择方法.

综上所述,本文的主要贡献如下:

(1)设计了无需枚举所有可能图实例的候选频繁模式预处理机制;

(2)将难于处理的不确定图转换为有效的确定模型进行处理,并通过算法后续工作恢复其不确定信息.

本文第 2 节综述相关研究工作;第 3 节定义不确定图数据集上的极大频繁模式问题;第 4 节提出基于随机游走的 K 极大频繁模式挖掘算法 RAKING;第 5 节给出实验结果;第 6 节总结本文工作.

2 相关工作

在图的频繁模式挖掘工作中,确定图上的 AGM[1]、FSG[2]及 Yan 提出的 gSpan[3]等都是针对频繁集进行挖掘的方法.因以上方法得到的结果集往往过大且不具有典型性,人们已经逐渐开始关注有约束的频繁模式挖掘算法:Tian[4]及 Chen[5]提出在宏观上对原有数据进行聚类或抽取,减小搜索空间后,再进行频繁集挖掘的算法;针对具体的约束,CloseGraph[6]是基于 gSpan 编码挖掘闭模式的有效算法;FOGGER[7]提出了频繁的"生成子子图"概念,挖掘频繁集的一个子集;近期 Yan 的 Leap[8]和 SigGraph[9]都是针对符合用户定义 p_value 的显著模式进行挖掘的算法.

在确定图上针对极大频繁模式的有关工作中，SPIN[10]首先从图数据集中挖掘出全部的频繁树，然后再通过频繁树重新构建出各个极大频繁模式；Margin[11]先将图数据组织成格，在搜索的同时不断对搜索空间进行剪裁以减少子图同构的计算，从而更易获得极大频繁模式.但是，因为不确定图的频繁子树也是不确定的，而且不确定图蕴含的全部确定子图空间巨大，即使进行一定的剪裁也很难有效枚举，所以这两种方法都不可以直接应用于不确定图.

随机化的算法因可在大规模数据上高效执行被广泛应用.在确定图上，ORIGAMI[12]通过随机化方法解决了获得有代表性模式的问题，但其输出不具有一致性，多次迭代后结果中仍可能漏掉一些重要模式.MUSK[13]方法则通过随机游走获得极大频繁模式集.近期，Hasan 在原有工作基础上提出利用随机游走对各类带约束模式进行挖掘的通用方法[14]，但以上方法对确定图进行的处理并没有考虑边及点的不确定性，不能很好适用于不确定图.

对于不确定数据的研究近年也已有了很多成果，如对不确定数据建模及管理的工作[15-16]，文献[17]介绍了最新不确定数据的相关技术，但这些研究仍然主要面向传统数据项.针对不确定图的研究才刚刚开始，其中已有计算不确定图中的最可靠子图[18-19]，对不确定图进行高效 TOP-K 查询[20]等课题.邹提出在不确定图上挖掘频繁模式的一些有效算法[21-23]：其中文献[21]首先将不确定图的蕴含子空间按边的大小组织为格的形式，然后在不断对空间进行搜索的同时根据 Apriori 性质对搜索空间进行剪枝，降低了子图同构次数，但由于其获得的是频繁集，所以结果数量巨大，可利用性低，且该方法基于枚举子空间，当数据量大时严重影响执行时间，不能快速完成.至今为止，在不确定图的研究方面仍未见针对极大频繁子模式的有关算法.本文综合考虑不确定图的概率模型与随机方法的自身特点，提出在不确定图上进行极大频繁模式挖掘的有效算法.

3　问题定义

可能世界模型被广泛用于不确定数据集的建模中，在该模型中各元组的任一合法组合均构成可能世界实例(instance)，实例的概率可通过相关元组的概率计算得到，可能世界实例的数量远远高于不确定数据集的规模.本文在可能世界模型的基础上对不确定图进行建模.

定义 1(不确定图).　五元组 $G=(V,E,\Sigma,l,p)$，其中 V 是无向图的顶点集，$E\subseteq V\times V$ 是边的集合，Σ 是标号集，l 是顶点标号分配函数，$p:E\to(0,1)$是边存在性的概率函数，是一个不确定图.

例 1.　多个不确定图共同构成不确定图集，如图 1 所示，在不确定图数据集中图结构不仅表示拓扑信息，图上的每条边也都有一个概率值表示该边的不确定概率信息；如图 1，G_1 的边$(A-A)$上权值为 0.8，表示该边以 0.8 概率存在.对于一个不确定图 G，若其边 e_1 的存在概率 $p(e_1)=1$，则表示边 e_1 一定存在.

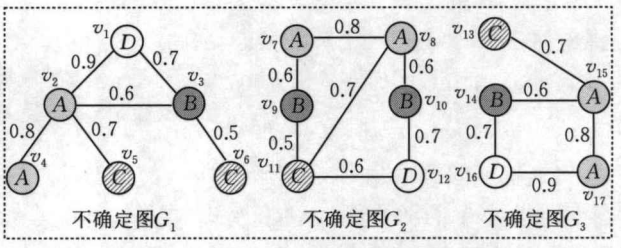


图 1　一个不确定图数据集 D 示例

对一个有$|E|$条边的不确定图，其所蕴含的确定子图数为 $2^{|E|}$.如图 2 所示为不确定图数据集中一个子模式 $M1$(见图 3)的 8 个蕴含子图($M1$ 有 3 条边，则其蕴含子图的 $2^3=8$ 个).

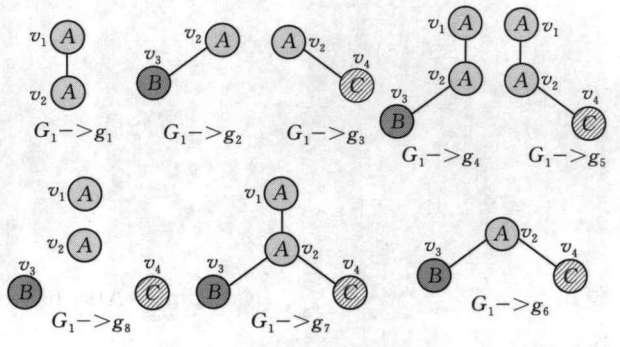


图 2　不确定图 $M1$ 的蕴含子图

定义 2(子图同构).　确定图 $g=(V,E,\Sigma,l,p)$与 $g'=(V',E',\Sigma',l',p')$子图同构，当且仅当其满足一个单射函数 $f:V\to V'$记为 $g\sqsubseteq g'$使得

$$\forall v\in V,\ l(v)=l'(f(v)) \tag{1}$$

$$\forall(u,v)\in E,\ (f(u),f(v))\in E' \tag{2}$$

如果 g 和 g'满足 $g\sqsubseteq g'$且$|V_g|\neq|V_{g'}|$，则称 g 真子图同构于 g'，记为 $g\sqsubset g'$.

确定图 g 是不确定图集 D 的子图模式，当且仅当 g 子图同构于 D 中至少一个不确定图 G 中蕴含的某个确定图.若 g 的边数为 k，所同构的 G 蕴含的确定图为 g'，则当 $g\sqsubset g'$，$|E_g|+1=|E_{g'}|$时，我们称 g'是 g 的直接超模式，g 是 g'的直接子模式.

定义 3(频繁子图模式). 对于不确定图的图集 D,用户指定最小支持度阈值 $minsup$,p_g 表示子图发生的概率,$sup(g,D)$ 表示确定图 g 在 D 中发生的期望支持度,使

$$\delta(g,p_g,G)=\begin{cases}p_g, & g\text{ 子图同构于图 }G\text{ 的概率}\\ 0, & g\text{ 非子图同构于图 }G\end{cases} \tag{3}$$

$$sup(g,D)=\sum_{G_i\in D}\delta(g,p_g,G_i) \tag{4}$$

设频繁模式集为 $\mathcal{F}$,若 $sup(g,D)\geq minsup$,则 $g\in\mathcal{F}$,g 是 $\mathcal{F}$ 的频繁子图模式.

定义 4(极大频繁模式). 设 F 为频繁模式集,$\mathcal{M}$ 为极大频繁模式集,称 p' 是一个极大频繁模式($p'\in M$)当且仅当

$$sup(p')>minsup,\ \forall p\sqsubset p',\ p\in F,\ p'\neq p \tag{5}$$

例 2. 如图 3 所示,$minsup=1$,$M1$,$M2$ 为不确定图集 D 中的两个极大频繁模式. 在 $M1$ 中,其支持度为各边概率之积:$sup(M1)=[p(E(A-A))\times p(E(A-B))\times p(E(A-C))]\times n=0.8\times0.6\times0.7\times3=1.008$. $sup(M1)>minsup$,但当为 $M1$ 添加任意一条边后其支持度都小于 $minsup$,则判定 $M1$ 为极大频繁模式,显然,其极大频繁模式的所有子模式都是频繁的.

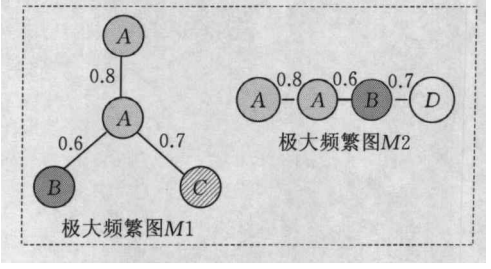


图 3 不确定图集 D 中的极大频繁子图

为了叙述方便,下文如无特殊说明所提图皆为不确定图. 在本文所建立的模型中,顶点和边都存在不确定性,但由于顶点的存在依赖于边,而边的属性则由边的两个顶点决定,所以本文所探讨的边不带标号. 这也与现实情况相符,在蛋白质网络,若已知一种分子结构时不能确定其分子特征,如果在结构的基础上确定了顶点的信息,则两顶点间的关系属性亦可确定,这是因为不同顶点上分子键的特性决定了边的属性.

本文研究在不确定图数据集中获得一致(各极大频繁模式以期望概率输出)的 K 极大频繁模式的挖掘问题. 具体定义为

输入:一个由大量不确定图构成的不确定图数据集 $D=\{G_1,G_2,\cdots,G_n\}$ 和期望支持度阈值 $minsup$,所需获得的极大频繁模式个数为 K.

输出:在不确定图中以一致概率输出所有极大频繁模式中的 K 个.

4 基于随机游走的图挖掘算法

算法主要由 3 个步骤构成:第 1 步将不确定意义下的图模型抽象为确定性空间,利用已有的确定性频繁图挖掘算法进行预处理,从而有效减小搜索空间;这一步骤使用的传统图挖掘算法与以往工作不同,传统的不确定图挖掘方法是先将不确定空间全部枚举,再在枚举出的确定空间内进行挖掘,本文则是在不考虑其不确定性的情况下进行挖掘,对不频繁出现的模式进行有效剪枝,再恢复其不确定性. 算法第 2 步将频繁子图模式恢复为不确定图然后使用剪枝后的模式空间构建成带有权值的随机游走空间. 第 3 步是在搜索空间上通过随机游走,不断检测所经过状态代表的模式,当获得 K 个极大频繁模式时,输出结果.

4.1 原始数据的预处理

4.1.1 在确定意义下进行数据预处理

定理 1. 在一个不确定图数据集中,所有不确定频繁模式是确定频繁模式的子集.

证明. 确定图是各边概率为 1 的不确定图,对于不确定图 g_1 其期望支持度为 $sup(g_1,D)=\sum_{G_i\in D}\delta(g_1,p_{g_1},G_i)$,若不考虑不确定性,等价于对于每一个模式 g,其支持度为 $\delta(g,1,G_i)$,p_{g_1} 为 g_1 的存在概率,由 $p_{g_1}\leq1$,显然 $\delta(g,p_g,G_i)\leq\delta(g,1,G_i)$,所以确定意义下获得的频繁图是不确定意义下的超集.

在本步骤中,将不确定空间抽象为确定的图集,使用对确定图进行挖掘的算法将支持度较低的模式剪枝. 本文采用当前较高效的 gSpan 算法并加以多种优化,在对原始数据进行预处理的执行过程同时记录各个频繁模式在确定意义下的支持度等信息.

4.1.2 恢复确定意义下的频繁模式的不确定性

本小节将恢复 4.1.1 节所得频繁模式的不确定性,并计算其在不确定意义下的期望支持度,对于模式 p,其期望支持度为各个模式存在概率与确定意义下支持度 num_{sup} 之积,即 $\prod_{l_i\in p}p_{l_i}\times num_{sup}$.

算法 1.　InitDataSet.

输入：不确定数据集 D，期望支持度阈值 $minsup$

输出：处理后带有期望支持度信息的频繁模式

1. 使用经优化的 gSpan 处理 D 中分离了不确定性的各个模式，并将确定意义下所得支持度等信息存储于各频繁模式；

2. 恢复各确定意义下频繁模式的不确定性，使用式(4)计算各个频繁模式在不确定意义下的期望支持度，若达不到不确定意义下的支持度阈值 $minsup$ 则剪枝；

3. 输出预处理后的结果集 D'.

算法 1 的时间复杂度分析：gSpan 算法时间复杂度为 $O(2^n \cdot 2^n)$，是指数级别的算法，但其保存了频繁子图 G 的子图同构列表，有效地减少了同构次数，更为重要的是，在本算法中此处的 n 仅是不确定图数据集抽取了不确定性后的数据规模，相较于不确定集概率空间内 n 的指数级的确定子图，其规模已减小了指数级别，设 n' 为所获得的频繁模式数目，在算法 1 其它步骤中所需的时间复杂度仅为遍历各个频繁子图的时间 $O(n')$.

4.2　构建 W_SAG 有效搜索空间

马尔可夫的平稳分布，是满足

$$\boldsymbol{\Pi}=\boldsymbol{\pi}\boldsymbol{P} \tag{6}$$

的概率分布 $\boldsymbol{\Pi}$，其中 $\boldsymbol{\pi}$ 是一个大小为 $|S|$ 的行向量，稳态分布是矩阵 $\boldsymbol{P}$ 以 1 为特征值的特征向量，如果一个链到达了平稳分布，那么它在所有未来时间都保持这个分布. 我们用 $\boldsymbol{\pi}(i)$ 表示向量的第 i 分量，若一个马尔可夫链是可逆的，则满足

$$\boldsymbol{\pi}(u)\boldsymbol{P}(u,v)=\boldsymbol{\pi}(v)\boldsymbol{P}(v,u),\ \forall u,v\in S \tag{7}$$

如果马尔可夫链的状态空间 S 是图 $G(V,E)$ 的顶点集，对于任意两个顶点 $u,v\in V$，$(u,v)\in E$ 意味着 $\boldsymbol{P}(u,v)=0$，这个过程称为在图 G 上的随机游走. RAKING 算法模拟一个有限的马尔可夫链(每个节点一个状态). 使用加权状态增长图 W_SAG (Weighted Statue Addition Graph)作为状态转移空间. 在 W_SAG 中的每一个点表示一个确定意义下的频繁模式，每个点都是被选做极大频繁模式的候选集(各点由 DFS 编码[3])，W_SAG 中的每条边表示频繁模式的一个可能的一条边的增长，图自顶向下构建，W_SAG 的顶部是被默认为频繁的 Φ 模式.

例 3.　如图 4 所示 W_SAG 图的顶部为空，向下依次为增长一条边后产生的模式，底部为经过剪枝后确定意义下的频繁模式，最终的 K 极大频繁模式就在 W_SAG 图中产生，底部的两个模式正是我们在图 1 中所示的不确定图数据集 D 的极大频繁模式.

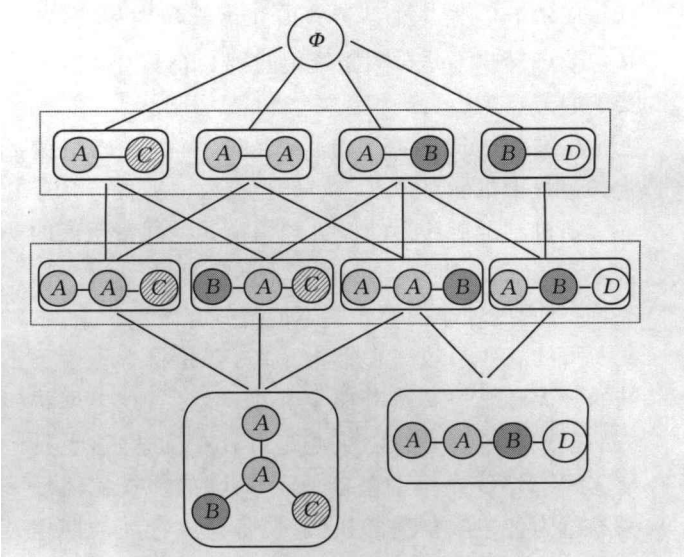


图 4　W_SAG 图原始模型示例

算法 2.　建立 W_SAG.

输入：算法 1 预处理后的输出集 D'，局部规模 m

输出：围绕当前节点构建 W_SAG 搜索空间

1. 在 D' 中随机选择一个频繁模式 p；

2. 从该模式出发，获得其全部的直接子模式及直接超模式；建立 p 与所有直接超模式和直接子模式的指针，构建 W_SAG；

3. 构建完成；输出 W_SAG.

算法 2 的时间复杂度分析：输入图数据集的规模为 n'，由于算法 2 在每一次执行时并不对 n' 的空间进行处理，而只是在 n' 集的局部进行用于随机游走 W_SAG 空间的构建，构建过程中按边的不断增长将各个模式编号且存储于一个类似格的空间中，并对各层间按图的边数排序，设各模式的平均大小为 $|p_n|$，则 p 将被分解为至多 MAX($|E|$)(最长边数) 层，每层的规模至多为 $|p_n|$，总的时间复杂度为 $O(|p_n|\times \mathrm{MAX}(|E|)\times \log|p_n|)$.

在 W_SAG 构建的过程中，本算法的一个重要的特点就是在局部而非全局构建 W_SAG，这将极大提高算法的效率，算法随机地选择一个预处理后的频繁模式，将以该模式作为当前中心点构建 W_SAG，当随机游走到达某无更下一层节点的状态时若判断其为极大频繁模式则输出，否则剪枝并回溯，继续游走. 对 W_SAG 的局部构建有效地减少了对全局数据进行处理的时间和空间消耗.

4.3　有效获得 K-极大频繁模式

定理 2.　W_SAG 具有 Apriori 性质.

证明.　由 W_SAG 的构建原理可知，自顶向下，下一层节点中的模式是其上一层与其有边相联节点的超模式，上一层节点是其有边相联节点的子模式，如果一个顶点所代表的模式是极大频繁模式，

则其所有上层节点都是频繁但非极大频繁模式;若某一节点模式不是极大频繁模式则必有其直接或间接下层节点为极大频繁模式,得证. 证毕.

根据 W_SAG 的 Apriori 性质,可以在随机游走的过程中不断对搜索空间进行有效的剪枝.

随机游走所获得的极大频繁模式往往由于度的大小不同而出现度较大的模式更容易被访问的情况,而其它度较小的模式则不能被等有效地输出,本文所指的高效一致获得极大子图,就是将各种度不同的极大频繁模式按其期望大小产生. 直观上度越大的模式就越容易在随机游走过程中到达,这也是建立 W_SAG 模型的基础. W_SAG 的加权性体现为通过为各边赋合适的权值,使各模式存在的期望与随机游走的转移概率相联系.

引理 1. RAKING 在 W_SAG 上的随机游走将收敛于平稳分布.

证明. 首先,频繁模式的数量是有限的,所以 W_SAG 是有限的;第二,由于 W_SAG 的底部是∅模式,图中任何一个模式都可以到达∅,所以对于 W_SAG 中任意两点 u,v 都至少存在一条可达的路径,可推出其是不可约的;最后,因为 W_SAG 是一个多阶段的模式增长图(每一层都是具有大小相同的模式),于是有可能会出现随机游走在同一个模式的各个子层循环游走,这时 W_SAG 就产生了周期性,但是这种周期性可以通过为图中每个顶点加入概率为 1/2 的自环而消除,此时图 W_SAG 必为遍历的. 由于任意有限、不可约且遍历的马尔可夫链都具有收敛于平稳分布的性质[24],所以,命题得证.

证毕.

对于图 $G(V,E)$,$|V|=n$,$w(a,b)\in W\geq 0$ 为边 $(a,b)\in E$ 的权值,我们假定 $w(a,b)=w(b,a)$,则转移概率为

$$P(u,v)=\begin{cases}1/2, & \text{如果 } u=v\\ \dfrac{w(u,v)}{\sum\limits_{x\in adj(u)}w(u,x)}, & \text{如果 } v\in adj(j)\\ 0, & \text{其它}\end{cases} \quad (8)$$

引理 2. 一个顶点的稳态概率与该点所有邻边权值之和成正比[13].

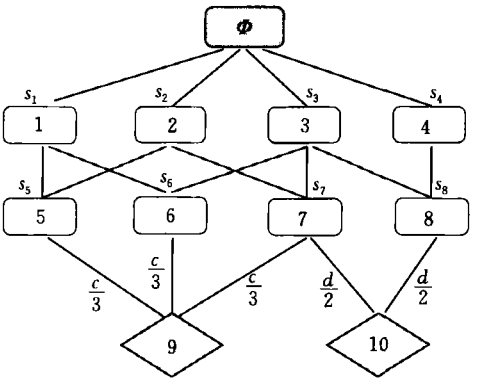
定义顶点的权:

$$w(u,v)=\begin{cases}1, & \text{如果 } u,v \text{ 都不是极大频繁模式}\\ p/d_x, & x\in(u,v)\text{中的极大频繁模式}\\ 0, & \text{其它情况}\end{cases} \quad (9)$$

式中 p 定义为极大频繁模式的期望支持度.

定理 3. RAKING 所提出的极大频繁模式生成方法使所有的极大频繁模式按其期望比例输出.

证明. 由引理 1 知 W_SAG 必收敛于平稳分布,引理 2 证明极大频繁模式的稳态概率与其相邻边权之和成比例,根据权的定义,对于极大频繁模式 v 其相邻边权值和为 $\sum\limits_{u\in adj(v)} d_u\times\frac{p}{d_u}=p$,设常数 p 是极大频繁模式的期望,命题得证. 证毕.



(1) 各边不加权时的稳态分布 π: (0.13,0.10,0.10,0.13,0.06,0.10,0.10,0.13,0.09,0.09,0.06)
(2)对于本文算法加权后进行随机游走的稳态分布 π:($c=3,d=3$) (0.10,0.07,0.07,0.10,0.05,0.09,0.09,0.17,0.12,0.15,0.15)
(3)对于本文算法加权后进行随机游走的稳态分布 π:($c=6,d=12$) (0.13,0.07,0.07,0.10,0.03,0.06,0.07,0.16,0.19,0.16,0.30)

图 5 W_SAG 图上的随机游走

例 4. 图 5 所示构建后的 W_SAG,计算其稳态分布时各个状态的稳态概率,(1)为各边不加权的情况,可以看到最后处于状态 9 和 10 的概率恰好是其度数的比 0.09/0.06=3/2;在(2)中设两个极大频繁模式的期望相同,使用式(9)计算可知:稳态时处于 9 和 10 的概率相等,在情况(3)中,设处于状态 9、10 两个极大频繁模式的期望分别为 $c=6$ 和 $d=12$,那么计算后可得稳态时处于两状态的概率分别为 0.16 和 0.30,当迭代次数增大时,趋近于 c 与 d 的比例.

算法 3. RAKING.

输入:不确定数据集 D,支持度阈值 $minsup$,所要获得的极大频繁模式个数 K

输出:K 极大频繁模式

1. 使用算法 1 进行预处理;
2. 使用算法 2 构建 W_SAG;
3. 按式(8)、(9)计算转移矩阵和各权值;
4. 在 W_SAG 上进行随机游走,并依 Apriori 性质,不断对搜索空间进行裁剪,若判断为极大频繁模式时输出,输出 K 次后算法结束.

算法 3 分析:RAKING 算法的主要运行时间用于预处理及构建 W_SAG,在随机游走过程中,相当

于对各个构建的 W_SAG 进行一次自顶向下的游动，其最大层数不会超过该模式的边数 $|E|$，属于常数时间，复杂度为 $O(2^n \cdot 2^n)+O(n')+O(|p_n|\times \mathrm{MAX}(|E|)\times\log|p_n|)$，总的时间复杂度为 $O(n^2)$.

RAKING 算法可以获得 K 极大频繁模式，此处参数 K 是原数据集中极大频繁模式个数的子集，当 K 值很大时，算法的执行时间会急剧增长，如果原数据集中仅有小于 K 个极大频繁模式，算法会在迭代多次后自动终止，在实验部分我们将讨论 K 值对结果的影响以及算法可支持的有效 K 值.

算法提出等概率输出各个极大频繁模式，是以文中定理 3 为基础，在算法执行过程中，我们在整个图集上以等概率选择一个模式进行扩展，构建随机游走的空间，这相当于在图集上等概率取得一个划分，由于划分后游走过程也是随机进行的，虽然可能出现划分大小不同的情况，但是对于每个极大频繁模式，其被产生的机会都是等同的. 在实验部分我们也会就所输出结果的质量进行讨论.

5　实验结果

我们利用实验考察本文算法的执行效率、不同数据规模对算法的影响以及随机算法所获结果的质量. 所有算法都在 STL 库支持下用 C++实现，Eclipse CDT 下 G++编译器通过. 用于实验的计算机具有 Intel Core 2 Duo1.66GHz CPU 和 2GB 内存，运行 Windows 7 操作系统旗舰版.

由于未见在不确定图数据集上进行极大频繁模式挖掘的相关算法，且在确定图上的挖掘算法与本文解决的问题缺乏可比性，本文主要是在不同规模及不同参数条件下验证理论分析的结果及算法执行的效率. RAKING 算法的实验数据集有两组，本文使用文献[3]中描述的两个分别拥有 340 和 422 个图的数据集，在进行不确定性处理后构成本实验使用的数据集空间. 不确定性处理的参数有 *average* 及 d，分别描述了所赋概率的平均值和各值在平均值上下浮动的范围.

不确定性处理过程如下：扫描数据集中所有不确定图，使用随机数产生器以 *average* 为均值，d 为浮动范围对每个图的每条边进行赋值，将不确定的概率值存入各个图的信息中.

实验 1.　首先考察在期望支持度阈值 *minsup* 变化的情况下 RAKING 算法的执行时间，取 $K=5$，使用两个数据集对算法执行的时间进行考察.

结果分析：从图 6 中可以看出当支持度增加后其执行速度有很大的提升，这是由于随着支持度的增加，其频繁模式的数量急剧减少，数据预处理的时间减少.

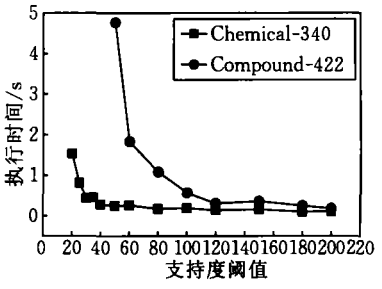


图 6　执行时间与支持度阈值关系

实验 2.　我们通过实验考察参数 K 值对算法执行时间的影响，我们取一个数据集为 Chemical340，$minsup=80$，调整 K 的取值，可得如图 7 所示结果.

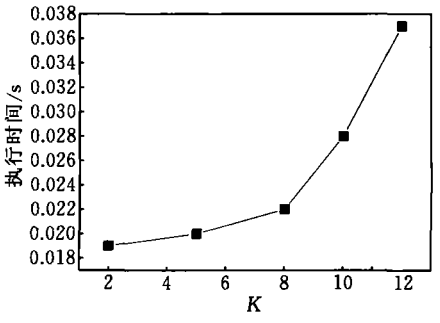


图 7　执行时间与参数 K 的关系

在这一部分实验中我们预处理后共获得 132 个确定意义下的频繁模式，恢复其不确定性数量虽然有所增加但其理想的极大频繁模式也只有 67 个，经过算法第 2、3 步骤处理后可以观察到，当 K 值增加到 8 以后，算法执行时间明显加长，这是因为算法随机选择一个频繁模式作为增长点，当 K 值增大时，迭代次数加多，执行时间加长. 但是我们统计 $K=2$，迭代 40 次后的输出结果发现，只发生了 4 次的重叠，这说明算法所获得极大频繁模式是分布均匀的，可以获得有效输出，根据实验我们得出，若图数据集中所有的极大频繁模式数为 m，当 $K<2m/3$ 时都能获得重复率较低、分布均匀的极大频繁模式集.

实验 3.　考察当 $minsup=50$，$K=5$ 时，两组数据中不同的输入规模对算法执行时间的影响.

结果分析：如果将第 1 步骤的时间如图 8(b)与整个算法执行的时间如图 8(a)在算法中分离，可知算法的多数时间花费在预处理的工作上，预处理后，不同规模上的数据获得 K 极大频繁模式的时间是相近的，这是由于算法使用局部构造及随机游走的方式，只需获得全局数据的部分信息即可得到正确结果.

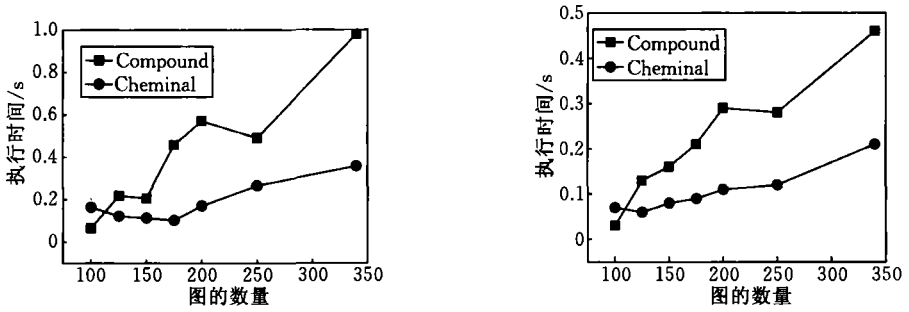


图 8 整个算法执行时间与图数量关系

6 结 论

本文研究了不确定图数据集上的 K 极大频繁模式挖掘问题,提出了不确定图数据集上基于随机游走的 K 极大频繁子模式挖掘算法 RAKING,RAKING 将随机游走与不确定模式的不确定性相结合,解决了不确定图蕴含子空间大且难于挖掘的问题.实验结果表明本文提出的算法能够高效地获得不确定图集合中的 K 极大频繁模式.

参 考 文 献

[1] Inokuchi A, Washio T, Motoda H. An apriori-based algorithm for mining frequent substructures from graph data// Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'00). Freiburg, Germany, 2000: 13-23

[2] Kuramochi M, Karypis G. Frequent subgraph discovery// Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM 2001). San Jose, California, USA, 2001: 313-320

[3] Yan X, Han J. gSpan: Graph-based substructure pattern mining//Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM 2002). Maebashi City, Japan, 2002: 721-724

[4] Tian Y, Hankins R A, Patel J M. Efficient aggregation for graph summarization//Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2008). Vancouver, BC, Canada, 2008: 567-580

[5] Chen Chen, Lin Cindy X, Fredrikson Matt, Christodorescu Mihai, Yan Xifeng, Han Jiawei. Mining graph patterns efficiently via randomized summaries//Proceedings of the VLDB09. Lyon, France, 2009: 742-753

[6] Yan X, Han J. Closegraph: Mining closed frequent graph patterns//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'). New York, NY, USA, 2003: 286-295

[7] Zeng Z, Wang J, Zhang J, Zhou L. FOGGER: An algorithm for graph generator discovery//Proceedings of the 12th International Conference on Extending Database Technology (EDBT 2009). Saint-Petersburg, Russia, 2009: 517-528

[8] Yan X, Cheng H, Han J, Yu P S. Mining significant graph patterns by scalable leap search//Proceedings of the SIGMOD 2008. Vancouver, BC, Canada, 2008: 433-444

[9] Ranu Sayan, Singh Ambuj K. GraphSig: A scalable approach to mining significant subgraphs in large graph databases//Proceedings of the IEEE 25th International Conference on Data Engineering (ICDE'09). Washington, DC, USA: IEEE Computer Society, 2009: 844-855

[10] Huan J, Wang W, Prins J, Yang J. SPIN: Mining maximal frequent subgraphs from graph databases//Proceedings of the SIGKDD. Seattle, WA, USA, 2004: 581-586

[11] Thomas L T, Valluri S R, Karlapalem K. Margin: Maximal frequent subgraph mining//Proceedings of the 6th International Conference on Data Mining (ICDM'06). Hong Kong, China, 2006: 1097-1101

[12] Hasan M, Chaoji V, Salem S, Besson J, Zaki M. ORIGAMI: Mining representative orthogonal graph patterns//Proceedings of the ICDM 2007. Omaha NE, USA, 2007: 153-162

[13] Hasan M A, Zaki M. Musk: Uniform sampling of k maximal patterns//Proceedings of the 2009 SIAM International Conference on Data Mining. John Ascuaga's Nugget. sparks, NEVADA, USA, 2009: 650-661

[14] Hasan M A, Zaki M. Output space sampling for graph patterns//Proceedings of the VLDB09. Lyon, France, 2009: 730-741

[15] Sarma Anish Das, Halevy Alon, Widom Jennifer. Working models for uncertain data//Proceedings of the 22th International Conference on Data Engineering (ICDE 2006). Atlanta, Georgia, USA, 2006: 7

[16] Sen P, Deshpande A. Representing and querying correlated tuples in probabilistic databases//Proceedings of the International Conference on Data Engineering (ICDE). Washington, DC, USA: IEEE Computer Society, 2007: 596-605

[17] Zhou Ao-Ying, Jin Che-Qing, Wang Guo-Ren, Li Jian-Zhong. A survey on t he management of uncertain data. Chinese Journal of Computers, 2009, 32(1): 1-16(in Chinese)
(周傲英,金澈清,王国仁,李建中. 不确定性数据管理技术研究综述. 计算机学报, 2009, 32(1): 1-16)

[18] Hintsanen P. The most reliable subgraph problem//Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases. Warsaw, 2007: 471-478

[19] Hintsanen P, Toivonen H. Finding reliable subgraphs from large probabilistic graphs. Data Mining and Knowledge Discovery, 2008, 17(1): 3-23

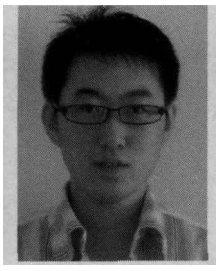
[20] Zhang Shuo, Gao Hong, Li Jian-Zhong, Zou Zhao-Nian. Efficient query processing on uncertain graph databases. Chinese Journal of Computers, 2009, 32(10): 2066-2079(in Chinese)
(张硕，高宏，李建中，邹兆年. 不确定图数据库中高效查询处理. 计算机学报，2009, 32(10): 2066-2079)

[21] Zou Zhaonian, Li Jianzhong, Gao Hong, Zhang Shuo. Mining frequent subgraph patterns from uncertain graph data. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(9): 1203-1218

[22] Zou Zhaonian, Li Jianzhong, Gao Hong, Zhang Shuo. Finding top-k maximal cliques in an uncertain graph//Proceedings of the ICDE 2010. Long Beach, California, USA, 2010: 649-652

[23] Zou Zhaonian, Li Jianzhong, Gao Hong. Discovering probabilistic frequent subgraphs over uncertain graph databases//Proceedings of the 16th ACM SIGKDD Conference of Knowledge Discovery and Data Mining (KDD'2010). Washington, DC, 2010, to appear

[24] Mitzenmacher M, Shi Daoji et al Translation. Probability and Computing: Randomized Algorithms and Probabilistic Analysis. Beijing: China Machine Press, 2007(in Chinese)
((美)米曾马克等著，史道济等译. 概率与计算. 北京：机械工业出版社，2007)



HAN Meng, born in 1987, M. S. candidate. His research interests focus on graph data mining.

ZHANG Wei, born in 1979, Ph. D., lecturer. His research interests include graph data management, wireless sensor networks.

LI Jian-Zhong, born in 1950, professor, Ph. D. supervisor. His research interests include parallel database, wireless sensor networks.

Background

This work is partially supported by the National Natural Science Foundation of China under grant No. 60903017, Heilongjiang University Student scientific and technological innovation grant.

In recent years, graphs have be used to model complicatedly structured data from a wide range of applications such as bioinformatics, pattern recognition, XML, communication network, chemistry, social network, World Wide Web, etc. Uncertainty is inherent in much of the data due to the imprecise characteristics of equipments or the nature of data. It is important and demanding to efficiently obtain the useful information from these database or sensor networks with uncertainty. Nowadays, the research on uncertain databases is one of the current research focuses. The existing work in the early stage focused on presenting data models for uncertain relational databases, the existing works have proposed algorithms for efficiently process queries on probabilistic graphs, and for mining frequent subgraph patterns from uncertain graphs. But for a maximal frequent especially in massive large graph database, there isn't any efficiently work. The paper proposes a random walk based K-maximal frequent pattern mining algorithm on uncertain graph set. Through RAKING K-maximal frequent patterns are selected from all maximal frequent patterns equiprobably. Theoretical analysis and experimental results show that the proposed algorithm can efficiently retrieve the K-maximal frequent patterns of an uncertain graph set.